PART I
CONCEPTS

Chapters

CHAPTER 1

Introduction to Research Data Management

1.1 What is Research Data Management?

Research data management is growing as a speciality because computerized data has become the basis for scientific research. The scientific method requires observation, inspiration, and discipline to produce believable results. A hundred years ago, a collection of small bound books was sufficient to record scientific observations. Scientists recorded information in scientific notebooks according to an analysis plan and carefully saved them after a project finished for potential future use. Modern research still depends on similar principles, but most scientists cannot rely on manual methods alone. The availability of computers encourages people to study increasingly large volumes of data.

The tremendous growth in the scale and number of research investigations has instigated a need for people who know how to manage research data. Computers have become standard tools in the last fifteen years. Regardless of the scope of a project, producing results depends on being able to access and manipulate information stored in computer files. The collection of files that must be managed can rapidly grow into a frustrating mess without a deliberate approach to research data management.

To better understand what we mean by research data management, let us establish a common terminology and put things in historical perspective.

Terminology

Five terms are fundamental to our discussion of research data management concepts. The first four intertwined definitions reflect our belief that effectively managing research data requires an engineering approach.

Research Data Management (RDM)

The systematic handling of information, ultimately stored in electronic form, to preserve the value of the contents for future scientific investigation or to answer specific research questions. Assuming the data accurately reflect reality, the value of research data depends on usability and timeliness. Excellent research data management will produce a practical framework for creating well-documented data of consistent quality, which permits research projects to proceed quickly and easily.

RDM System

A systematic set of research data management activities and procedures related to acquiring, manipulating, documenting, and storing research data in a computer environment. An RDM system is composed of interrelated RDM processes.

RDM Process

A set of activities or procedures directly related to accomplishing a specific research data management objective. Examples of RDM processes include data collection (asking survey questions), data entry (keying survey answers into a computer), data cleaning (identify and fix data problems), and backup (saving extra copies of data sets).

RDM Monitoring

A set of activities or procedures for tracking and evaluating an ongoing RDM process for management and control purposes. The goal of RDM monitoring is to quickly identify problems so that resolutions may be found and implemented in a timely fashion. Typical monitoring processes could involve computing data entry error rates, auditing a data update log, or reviewing program documentation.

Resource Profile

The resource profile of a research project defines the available resources and environmental restrictions. The most important components cover personnel, money, time, computing equipment, and technology tools. A resource profile represents how tangible and intangible assets are balanced against liabilities that impact the RDM system.

Research data management integrates related components—RDM processes and RDM monitoring—within the context of a resource profile. To manage a functional RDM system, an "RDM analyst" must be skilled in designing an RDM plan, setting up RDM processes, establishing effective monitoring, maintaining a working system, and correcting any unexpected problems. An experienced RDM analyst will be strongly influenced by the resource profile when choosing or designing an RDM process. For example, suppose hiring people would be less expensive than buying and supporting special computer equipment. The reduced labor cost could justify recommending a more labor-intensive RDM process over an automated solution. We discuss the "Influences of the Resource Profile" further in Section 1.4.

We explain more about RDM processes and RDM monitoring in later chapters. You will find ideas about planning common RDM processes in Section 2.2, "Tips for RDM Planning." Section 3.2 presents basic answers to the question "Why Monitor RDM Systems?", and Sections 9.3 and 10.3 include examples of monitoring processes.

The RDM Legacy

The evolution of research data management is influenced by two pressures: the habits of scientists or computer specialists and continuous improvements in the computer environments available to research projects. Let us review a little history before we present our philosophy of research data management.

Before computers became available for general research, scientists could manage their own scientific data without any special help. Maintaining a scientific notebook mainly required basic writing skills. For larger projects, a scientist could easily train assistants to help record observations or organize information. Neatness and patience were the management skills needed to control or improve the quality of the data. Research data management was not a concern.

Jump forward to a time when only cumbersome mainframe computers were used to process and analyze research data. Although most projects relied on trained computer specialists or technical staff to create and analyze data, scientists still used a scientific notebook as the method for keeping track of the data and associated analysis procedures or results. Managing the research data for a small project could just mean handling a box of punched cards carefully to keep it safe from ruinous forces. ("Don't drop that sorted box of cards.") Certain computer facilities automatically made backup copies of data files and generally helped assure that research data were protected. Research data management was in its infancy.

The introduction of affordable and powerful microcomputers has dramatically changed the data management responsibilities of a researcher. Research data may span hundreds of computer files, include thousands of variables, and have hundreds of thousands of observations. Electronic files now may contain data values, results, documentation, or even scanned images. No one can manage such a large amount of information using a scientific notebook approach. Even a small project rarely has only a few data files. Understanding the variety of possibilities for managing research data is difficult without previous experience. However, the scientist is ultimately responsible for the quality of the data used to construct research results. Research data management is now a necessary part of any project.

1.2 A Philosophy for Research Data Management

Our philosophy for research data management recognizes that scientific research varies widely in complexity and purpose. However, for any project you can always:

- Strive to make your RDM system as simple as possible.
- Create the simplest, most practical RDM processes that will use your project's resource profile effectively.
- Work creatively to accomplish RDM objectives with the resources that are readily available.
- Make continual improvement of your RDM system a high priority.

Research data management requires a team approach, but strong or understanding leadership helps immensely. Use a sequential approach initially, but stay flexible as the situation evolves. To create the most useful research data, the project team should:

1 plan and document the RDM system
2 implement and document RDM processes
3 monitor, back up, and document RDM activites
4 revise processes and document revisions
5 archive and document project completion activities.

The legacy of the self-sufficient scientist influences how people perceive research data management. We believe that increasing reliance on "user-friendly" computing technology means that people overlook the discipline required. An inexperienced researcher may assume that data management is a clerical task requiring little training. However, keeping a valid scientific notebook required a disciplined mind and early computers were supported by people with engineering skills. Bear in mind that disastrous data management failures could lead to scientific misconduct cases or product liability cases.

You can create the best RDM systems by always choosing the simplest alternative, especially as a system evolves. Take the time to fully exploit all components of your project's resource profile. Make informed decisions that respect the capabilities of the project team and the computing environment. You may be surprised by the elegance and sophistication that results from following this philosophy.

1.3 Essential RDM Issues

In this section, we introduce the fundamental research data management issues of planning (or lack of planning), documentation, protection, and dealing with an imperfect RDM system. Data security procedures, monitoring, and team support represent different approaches for protecting research data.

If you are unfamiliar with the terminology, please have patience. You will find more explanations in later chapters. Chapter 2 covers planning. Chapters 3 and 4 include guidelines for documentation and protection techniques. We present strategies for dealing with common RDM problems in Chapter 5.

Planning

RDM planning should cover all aspects of the RDM system. Important areas include, but are not limited to

- data acquisition
- data verification and validation
- data manipulation and analysis
- result reporting
- back up and archiving
- monitoring procedures
- documentation procedures and standards
- teamwork (training, communication, support).

Comprehensive planning produces the best data. Resist the pressure to forge ahead without an RDM plan due to a perceived lack of resources or time. You must plan every major RDM process if you want a smooth, cost-effective project. We present an "Overview of RDM Process Design" in Section 2.1.

Three areas that should be emphasized during planning are monitoring data acquisition, reducing the number of data transitions, and designing database structures for analysis. Imagine the hassles if part of a major data collection effort had to be repeated because a data acquisition process was poorly monitored. An unnecessary data transition increases costs and becomes a source of potential data errors. For example, transcription of answers from one form to another to change values to coded numeric values could introduce errors.

Research databases should be structured with specific analyses in mind. For instance, if an analysis requires information from multiple sources to be linked, then the data structure must include appropriate identifiers in all files to link records. Note that a multi-disciplinary approach to planning database structures will produce optimal database designs in the long run.

Documentation

The scientific notebook still plays a role in research data management in the sense that documentation is the heart of an RDM system. No documentation system is ever perfect, but poor documentation beats no documentation anytime. Undocumented knowledge about research data fades with time. Obvious details unrecorded today can become challenging puzzles six months into the future. An undocumented system cannot be validated, audited, or used as a guide for future work. In the end, the history of the research data has a bearing on the interpretation and credibility of analysis results. See Section 3.1, "What Standards are Required?", and Section 3.3, "Why Is Documentation So Important?", for more on documentation.

Documentation is an important mechanism to make sure people communicate frequently and clearly. Communication between team members helps preserve the quality and value of the data. Think of documentation as the means by which work can continue when a team loses a member.

Protection of All Kinds

Data protection can be defined many ways. Data security procedures protect data against corruption, physical loss, or unauthorized access. Monitoring protects the value of research data by identifying potential problems and keeping the RDM system running smoothly. Indirectly, fostering team support for the RDM plan is the best protection for achieving RDM goals.

Threats to data security are easily understood but are still often ignored. Equipment failure, a computer virus, or a natural disaster (fire, flood, earthquake) can quickly destroy data files and documentation. Other threats may necessitate access restrictions. For many research situations, precautions are required to maintain confidentiality. One possibility is that lawyers or competing companies must be kept away from confidential data or results. Ignoring any of these threats is an invitation for awkward and unwanted inquiries. See section 4.3, "Keeping Data Safe," for data security suggestions.

RDM monitoring protects research data by detecting problems early, which decreases the risk that data problems will overwhelm a project. Suppose data acquisition that required logistically complex field work occurred without any monitoring and later data errors were found. The remainder of the project might be aborted because the cost of rework would be too great. Think of monitoring as protection against hidden weaknesses in the data or the RDM system.

Ultimately, the value of research data is protected best when the project has strong leadership and team support for RDM goals. A high level of quality is most easily attained when everyone understands the importance of data quality, the need for documentation, and the reasons behind following RDM standards and procedures. Obviously, to create such a situation requires both "business" management and research data management skills.

Imperfect RDM Systems

Although this book stresses a proactive approach to data management, we recognize that inheriting a poorly designed RDM system happens often. Rest assured that improvement or recovery from minor failures is frequently achievable once problems are identified. Refer to Chapter 5 for ways to deal with "Common Research Data Management Problems." Remember that everyone involved wants to learn how to correct that system because, presumably, they care about the results.

One reason imperfect RDM systems exist is that during project planning many scientists focus on the complexities of their specialty to the exclusion of other issues. Data management is overlooked and when this oversight comes to light, the response is often "I didn't know it mattered." Even experienced researchers may initially ignore data management issues, perhaps because they assume new technology has simplified RDM processes. They may not realize that dedicated staff are producing usable data from a deficient RDM system only after continuous struggles. The good news is that serious researchers will listen when presented with a thoughtful plan for improving data quality or RDM system efficiency.

1.4 Influences of the Resource Profile

To make smart decisions about data management strategy, you must completely understand the resource profile of your project. Think of the components of the resource profile as positive or negative attributes that interact. After general comments, we discuss the impact of the most important components: work environment, personnel, finances, timeframe, and technology environment. The case studies in Part II each represent different resource profiles and are compared in Section 6.2, "Comparing Case Studies."

The resource profile represents a balance sheet of assets and liabilities, which can be difficult to define thoroughly. For a "one person project," a one page inventory of tangible assets (paper, disk storage, laboratory equipment, etc.) and intangible assets (research experience, commitment to success) could be balanced against perceived liabilities (small budget, slow computer, conflicting responsibilities). For larger projects, producing a comprehensive balance sheet may require an experienced RDM analyst or "project director." Intangible items are especially difficult to understand. In any case, the effort to describe the resource profile is necessary to avoid wasting time making uninformed RDM decisions.

Perceptions of cost-effectiveness are influenced by the resource profile. For example, if people consider automated data entry (optical character recognition, touch-tone phone) too expensive for a small study, then manual data entry will be used. However, for a large project with a different resource profile, using optical mark readers could be approved because people perceived that the automated process would be cheaper when compared to manual data entry. When making RDM decisions, an experienced RDM analyst will be influenced more by the resource profile than by the experimental design of a study.

Research Workplace

The research workplace can provide assets or restrictions. For instance, people working at an oceanographic research organization with easy access to the water would count the work environment as an asset. A vaccine researcher working in a delta region infested with cholera might be more equivocal. The watery delta would be a hazard to the research staff and might make hiring difficult. It is not accidental that major research centers are often near urban areas, which provide ready access to extensive libraries, computing facilities, and potential research staff.

Personnel

The availability, or lack of, personnel (human resources) can drastically alter RDM planning decisions. With enough time you can train people and develop a team with the necessary skills. When time is short, you are restricted to whoever is immediately available. Hiring experienced staff is costly and may take time. Thus, whether or not personnel is an asset or a liability depends on the level of financial support or time available.

A special aspect of this component is whether or not people in leadership positions appreciate the importance of RDM goals. If decision makers are not expected to approve hiring for RDM activities, personnel may be considered part of the intangible liabilities. Having a project leader who has experience from similar projects would be an asset.

Finances

Money is a liability when your budget is too small; otherwise it is an asset. When money is short, other necessary items are usually in limited supply. A "make do" attitude is the order of the day. In the rare cases where financial support is more than adequate, do not forget the intangible cost of the time spent waiting for delivery of equipment and materials.

In this book, we assume that everything can be valued in monetary units. Putting a value on tangible items is easier because purchase prices or depreciated values are known. Intangibles such as goodwill or risks are harder to value. The intangible cost of risks contribute significantly to total cost of an RDM strategy. You want to answer the question: How much will it cost if a bad event were to occur? For example, suppose the bad event were the failure of a hard drive. You could estimate the cost by multiplying the probability that a hard drive fails by the estimated cost of a worst case scenario. Estimated costs would cover replacement parts along with labor costs for hardware repairs and restoring all data files on the drive.

One possibility is to estimate the cost of a risk as the average probability of the risk occurring multiplied by the predicated total cost of the worst case scenario. This type of conservative estimate places a higher value on the risk of a bad outcome and tends to encourage careful planning.

Timeframe

The timeframe of a project may be viewed in the same way as financial support; too little time is a liability. In contrast to money, the distribution of time is uniform. The clock ticks the same for everyone. Time can be associated with intangible costs. For example, suppose a competitor will reach the marketplace with a new product first if your project is delayed. Most project teams accomplish more in less time with more money, but the relationship is not linear. Money is less important if working slowly with inefficient tools and older technology is acceptable.

Technology Environment

The level of technology available for a project must be factored in if you want to have practical RDM processes. This applies not only to computers but also to the associated standard operating procedures and any other helpful technology.

Consider three different computing environments. In the first, electricity is limited and portable computers are the only practical option. In the second, powerful microcomputers are available but the disk space is not sufficient to store large intermediate work files. For the third, "free" computing power is provided by a computer network and a mainframe computer, but access is restricted to specific evening hours. Which environment would be an asset if you were setting up a project that had ten large (>100,000 observations) data files? Which would be an asset if the project had ten small (<500 records) files and must be finished in two weeks?

The sophistication level of computer-related standard operating procedures may affect RDM choices. Building on existing standards is usually easier than breaking new ground.

Because an RDM system includes procedures for documentation and communication, many other types of "technology tools" can be assets or liabilities for a project. Think about how the existence of telephones, facsimile machines, copiers, or electronic mail might impact proposed RDM monitoring.

1.5 The Real World

We find research data management a fascinating discipline because every project presents unique opportunities for creative thinking. One secret to making practical decisions is to recognize the limitations of the real world.

Technology is wonderful, but respect the strengths and weaknesses that differentiate people from machines. Relying on machines is best for repetitive tasks requiring exacting precision, while relying on people is best for tasks requiring judgment and creativity. Many tasks fall between these two extremes, and the correct decision is not necessarily obvious. For instance, suppose an RDM system will include a series of repetitive tasks that last only a short time. Although these tasks are repetitive, each is different. People could be trained faster than machines and will do repetitive tasks properly for a short period. In such a situation, automation may be too troublesome and unjustified.

A truism that applies to research projects is called the outcome triangle. The three sides are cost, quality, and speed. You can shorten any two sides but never all three at the same time. For instance, a low cost project that has the luxury of ample time can still produce high quality research data. But data quality will be compromised if timelines are shortened and the budget remains low.

Try to avoid the tendency to view the "fantastic" capabilities of computing equipment as magic. ("Let the computer handle it.") Perhaps using computer technology is magical, but first someone must understand an RDM objective and provide the proper instructions. You must design a system that fits your project's resource profile to benefit most from "computing magic."

Chapter 1: Tips To Remember

- *Be practical, pick the simpler solution*
- *Documentation is the heart of an RDM system*
- *Good science depends on planning for research data management, not "computer magic"*
- *The resource profile impacts RDM decisions*