Demystifying the Academic Research Enterprise

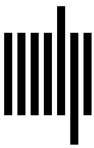
Becoming a Successful Scholar in a Complex and Competitive Environment

By: Kelvin K. Droegemeier

Citation:

Demystifying the Academic Research Enterprise: Becoming a Successful Scholar in a Complex and Competitive Environment
By: Kelvin K. Droegemeier
DOI: 10.7551/mitpress/14740.001.0001
ISBN (electronic): 9780262377201
Publisher: The MIT Press
Published: 2023

The open access edition of this book was made possible by generous funding and support from The MIT Press Frank Urbanowski Memorial Fund



The MIT Press

5

Becoming a Detective: Finding What You Need and Using It Effectively

Chapter Overview and Learning Objectives

Similar to a detective, researchers must find appropriate data and other resources to inform their work. Yet, simply obtaining or creating data is only part of the story. This chapter describes the information and evidence-gathering process, from becoming familiar with previous work to identifying sources of information needed (and whether they already exist or have to be created) to validating source material and ensuring its quality and appropriateness for use. It also highlights data analysis, synthesis, and visualization as tools for discovery and understanding. After reading this chapter, you should

- Understand the differences among various sources of data/information for performing research and creative activity, and the importance of becoming familiar with such sources in the context of your own work;
- Be able to explain the differences between primary and secondary sources;
- Understand the importance of source validation, quality assurance and quality control and the differences among them; and
- Have a general understanding of information synthesis and analysis and the different approaches used by various disciplines.

5.1 Becoming Familiar with Previous Work

When my sister and I were little, we loved to pretend we were detectives, searching for clues that would solve some mysterious crime. If you are a fan of detective movies or mysteries, you will love this chapter! Why? Because one of the first things you need to do, after identifying an interesting idea to explore, or a fascinating question you feel needs to be answered, is to piece together facts about work that has been done previously. Often this is referred to as conducting a literature review or landscape analysis, though in contrast

to the work of real detectives, the information usually is readily available. In fact, a survey of the literature is *the* first step in research and creative activity, and also in writing a grant proposal (chapter 6), because it accomplishes a number of things.

First and foremost, studying previous work lets you know the extent to which the idea you wish to explore, or the question you wish to answer, already has been studied. It also provides a context for your own work, helps identify gaps in existing knowledge that you may be able to fill, teaches you about previous approaches and techniques as well as information and data sources, and gives you a broad perspective of how knowledge and understanding regarding a particular topic have evolved over time.

Numerous sources exist for building your understanding about previous work, and the most important sources for many disciplines are refereed or so-called archive publications. They bear these names because they have been subjected to rigorous peer and editorial review and are part of the world's repository or archive of our current state of understanding. As discussed below and more extensively in chapter 7, such review does not always guarantee correctness, but rather evaluation and scrutiny by experts which, for the most part, does ensure that mostly original, high-quality scholarship is published. That said, the sources for understanding previous work do not end there. Numerous others exist, including books, oral histories, personal views of researchers you may obtain through conversations, opinion pieces, review articles, conference proceedings, video and audio recordings, diaries, paintings, sculptures, and scripts.

As we all know quite well, accessing information today is relatively easy owing to the Internet. Many previously unavailable documents, images, and audio and visual media are now available at the click of a button. Indeed, digital libraries abound, and open access frameworks (section 11.2) now make scholarly publications widely available (e.g., PubMed Central; http://ncbi.nlm.nih.gov/pmc). However, in our new highly connected, instant-access world, one can quickly become overwhelmed with the sheer volume of information available. Consequently, new tools involving artificial intelligence are becoming widely available to synthesize hundreds to thousands of publications in a matter of seconds (e.g., the CORD-19 data base, described in section 11.2) And, as noted later in this chapter, care must be taken within this universe of material to ensure that sources are trustworthy and of high quality.

How does one deal with all of these challenges?

Most formal publications, such as journal articles and monographs, contain an abstract, an introduction, a description of methods used, findings and an explanation of them, and conclusions, which sometimes include comments

about future work. You could literally spend months or years examining all materials relevant to your research topic or question, so a good strategy in the case of journal articles is to begin with a few you feel are most relevant. Read them entirely. You then will quickly realize these papers cite numerous other papers in their bibliographies. All of a sudden, you now have dozens to hundreds of related papers to explore. You may wish to read some of them front to back, but for others, it is best to read only the abstract and conclusions, diving into the rest of the text if the article is especially relevant. This allows you to examine a large body of scholarship in a reasonable amount of time, with deep dives into certain papers that are more relevant to your topic than others.

By virtue of this process, you also will come to know pioneers in the field and which works are viewed as seminal, especially via use of tools such as the h-index (Hirsch 2005) and its variants. Meta-analyses and review articles are excellent places to begin, if they exist for your chosen topic, because they summarize numerous studies over time and contain extensive bibliographies.

One important benefit of studying previous work is setting a context for your own. That context helps both you and others understand which knowledge gaps exist and how your work will help fill them. This contextualization may seem obvious, but I cannot begin to count how many times I have read a grant proposal in which the investigator jumps immediately into their own idea without setting the stage for me. As a result, unless I have deep expertise on the topic at hand, I have little idea whether their approach is novel, if the questions being asked have already been answered, or if their potential contribution will be valuable.

From a practical point of view, when performing literature reviews or other studies of previous scholarly work, you will find it helpful to make notes, including questions and comments. Numerous apps and programs exist to do so, and you can then easily cross-reference and index the material. Knowing the authors, composers, or artists, and the date or dates of the work performed, are especially important and help you recall a particular study when needed, such as in preparing for a thesis defense or a seminar.

One final point regarding becoming familiar with previous work. Unfortunately, published papers and monographs, recordings of performances, or videos of productions show only the end product and not the often-circuitous and difficult path taken to get there—a path that can be fraught with frustrations, dead ends and restarts. Do not simply believe that what you read is the final story, or that a particular piece of music, once played, can never be interpreted by being played in a slightly different manner! Respect previous work, but do not be afraid to challenge it, improve upon it, or use it as a launching pad for a new idea.

5.2 Assessing Your Need, Identifying Sources, and Collecting/Protecting Resources

Once you have completed your review of previous work on the topic you wish to study, and have developed a hypothesis, strategy, and set of tools or procedures for studying it (chapter 4), it is quite likely you will need data, artifacts, records, or other source materials in order to proceed. This is similar to the work of a detective, who needs to gather evidence. Knowing *what* you need may not seem obvious for research, especially if the topic being studied is particularly complex. However, this is where the hypothesis plays a key role because it helps frame, and in fact sets boundaries for, the sorts of resources you likely need. You will see specifically how this works in the exercises associated with this chapter. Not surprisingly, as one proceeds through the research process, the need for additional resources sometimes arises, though how and whether such additions can be used depends upon the topic and research methodology being used. When considering resources, they either exist or they don't. You either go find them or have to create them.

For the situation in which the resources you need already exist and simply need to be obtained, you must determine whether and how you can access then. Such resources typically are divided into two broad categories: primary sources and secondary sources.

A primary source is one that is original or represents a firsthand account and is purely factual. Examples of primary sources include ancient manuscripts in public or private archives, diaries, original or raw data sets from environmental or space observations, data from surveys of people, historical records such as presidential papers and audio recordings, output from computational models, or art objects. Secondary sources, on the other hand, are descriptions, interpretations, or evaluations of primary sources. In other words, they are not firsthand accounts but rather assessments. Examples include articles that review the original work of another person, newspaper and magazine articles, and opinion pieces.

If it turns out that what you need does *not* already exist, then you must collect or create it, and in so doing make a new primary source. For example, you may need to design a survey to evaluate public attitudes toward a particularly important and sensitive issue, such as gun control, or collect water samples from a lake that is infested with a rare form of algae to understand how it became toxic. You may need to create a computer model of the atmosphere on Mars and run simulations to understand how it has changed over the past several million years. You may need to interview a famous artist to understand their creative process and how they have passed it along to protégées. Or you

might even need to observe small children in various classroom settings to understand how their uptake of information differs from teacher to teacher or school to school. Some of the information collected by you may be confidential, so you may need to protect the identities of those who provided it, such as in certain types of surveys or clinical trials for experimental drugs. Protocols for such activities are discussed in chapter 10.

Irrespective of the sources and information used for your study, it is extremely important to protect it in other ways as well. If you are dealing with primary sources that you are not allowed to physically possess, then any notes you take, or copies or photos you make of the sources, need to be copied, backed up, and stored in one or more safe locations. This also is true of any and all notes you make or other information you collect as your research proceeds, say in the form of a laboratory notebook,[1] for the following reasons.

First, natural disasters, such as floods or fire, can quickly wipe out massive amounts of work in just a few minutes. Nothing is more devastating to a researcher than to lose critical information that either cannot be recreated or is expensive to recreate. This is why your saved versions should be stored in multiple locations, especially different from where you perform your work. For example, regularly scan your laboratory notebook or daily work summary and save it on a thumb drive in your home as well as office, or in the cloud. Make multiple copies of digital data and do the same with them.

Second, having a complete, archived record of your work—something often referred to as a workflow—is important when organizing your findings for publication or presentation because it is easy to forget seemingly minor but important details. Third, as discussed in section 4.7, the ability to reproduce research results depends upon knowing the steps involved in the process, so it is important to have multiple copies of such records. Fourth, as discussed in chapter 9, if for some reason the ethics and integrity of your work are questioned, having a complete record, with redundant copies, will position you to address concerns raised.

Finally, new open access frameworks (section 11.2) often require that you provide public access to information used in creating a publication. Having backup copies ensures that you can meet this requirement, which allows others to build upon your work as you have built upon theirs.

5.3 Source Validation, Quality Assurance, and Control

If you were a real detective trying to solve a crime, you would want to make sure the evidence you gather is factual before charges are brought and the trial begins. Likewise, regardless of whether you identify and gather information

from existing sources, or create completely new information, the associated validity and quality of this information are exceptionally important and must be established at the outset, before research begins. Methods exist, in every area of research and creative activity, for information validation, quality control, and quality assurance, though terminology varies among them. Because space does not permit a comprehensive treatment of validation, particularly because of variations across disciplines, I focus here on the basics. Additional resources may be found in the references (e.g., see Arthur 2017 and US Geological Survey n.d.-b).

In the context of research sources, validation concerns the confirmation or substantiation of the source. As noted in chapter 4, in historical research as well as arts and fine arts scholarship, validation involves addressing issues such as when and where the source was produced, by whom, the reliance of the source upon earlier material, the original form of the source, and the credibility of source contents. In the case of computer codes, such as models of physical, biological, and other phenomena, validation involves various tests to determine whether the code is performing as designed. Note this has nothing to do with the realism of the results produced by the computer model, but rather only that the code was properly constructed based upon its underlying mathematical framework and is working as intended.

Physical instruments used to observe the natural world also undergo validation as well as calibration procedures (sometimes referred to as cal/val). For anything in the digital realm, in this day of computer hacking and other nefarious intrusions, it is especially important to confirm the integrity of sources.

The terms "quality control" (QC) and "quality assurance" (QA) are sometimes used interchangeably, especially in the context of collecting new research data or processing existing data. In fact, QC and QA are quite different. Before examining them, we need to be clear about our use of the term "data," because data exist in every discipline and are foundational to research and creative activity.

Although data can be an elusive concept, I will define them as follows (and note—the term "data" is always considered plural): *Information that provides a quantitative and/or qualitative description or characterization*. This definition is equally applicable to humans, the atmosphere, core samples of ice collected in the Antarctic, an orchestral performance, and a piece of art. Data are not things. Rather, data are descriptive information. And I should mention that output from a computer model, such as a weather forecast model, is called just that—output. The term "data" typically is reserved for information associated with real phenomena, objects, processes, or activities, not those simulated with a computer.

With that preface, the phrase "data quality" is a general term encompassing attributes, both qualitative and quantitative, that characterize a particular data set. Data are said to be of high quality if they accurately represent the phenomenon or state of a system they were intended to portray and are appropriate for the intended use. In that regard, data quality may be subject to interpretation, and because of this, specific values are assigned to quantitative measures, and descriptions assigned to qualitative measures, in specifying data quality.

Data QA represents criteria and processes utilized to prevent problems occurring with data, say as they are being collected. Consequently, QA often is referred to as *defect prevention* because it proactively seeks to ensure that data meet appropriate and agreed-upon quality standards for the problem at hand. Examples of QA in practice include calibrating and siting instruments, say to measure air temperature and wind, so as to avoid false readings caused by nearby buildings and trees, as well as survey questions worded in a neutral manner so as to avoid evoking a desired response.

In contrast to data QA, data QC is the process by which data are subjected to various processes, following collection, to determine whether they meet stated quality goals. For this reason, QC often is referred to as *defect detection*. Examples include the use of automated algorithms to detect missing data, outliers, or suspicious values, and a variety of errors, such as those arising from a faulty measuring instrument (which could be physical device or a survey given to people). QC also addresses whether a given data element is representative of values nearby in space and time, assuming such should be the case.

In some cases, the QC process may determine that, although a data set is quality assured, sufficient issues exist to preclude its use for a particular purpose. That is not to say the QA'd data are unusable if they fail the QC test. In fact, a quality assured data set may be perfectly suited for one application but not for another. Consequently, it is important for you, as a researcher, to know the difference between QA and QC, and to make sure you have a complete understanding of your data or source information, including quality, before using them.

A final but very important point is that, to the extent possible, original or raw data should *never* be destroyed because QA and QC procedures continue to improve with time. A great example, in the context of climate change, is the global temperature record, to which various corrections need to be made to account for how instrument characteristics and placement have changed over decades.

5.4 Analysis and Synthesis

The analysis of data and information are foundational to research and creative activity, more so now than ever before. In fact, thirty to forty years ago, many disciplines were largely experimental in nature, including in the biological and chemical sciences. Today, those same disciplines, though still utilizing theory and experiment, are among the largest generators and users of digital data, ranging from high-throughput genomic sequencing to computer-based molecular modeling to three-dimensional imagery of ancient relics. Likewise, explosive growth in digital humanities, the ability of social and behavioral scientists to capture and study vast amounts of information from online surveys and social media, and the ability of atmospheric scientists to model, with exceptional realism, the entire earth system, have completely transformed the research enterprise. Coupled with open access frameworks (section 11.2) that make vast amounts of data available to anyone at the click of a button, our world—and to be certain, the research enterprise—is literally drowning in data.

Extraordinary potential exists in this sea of information to study entire problems, rather than breaking them into simpler parts (the so-called reductionist method), and to bring to bear all of the relevant disciplines needed to do so (chapter 13). This requires new approaches in data analysis, so much so that specialists in data informatics and analytics, and new tools such as artificial intelligence and machine learning, now dominate the research landscape. In other words, like a detective, you need to bring all of the pieces of evidence together into a coherent picture that tells a story!

As was the case for other topics we have discussed, the general topic of data analysis has its own lexicon, the terms within which differ in their meaning depending upon the discipline or nature of the application. You will find terms such as data synthesis, assimilation, and fusion used to describe the bringing together of data—perhaps of a similar or different type—in a manner that broadens and/or deepens understanding beyond that which otherwise would occur. This is a difficult but particularly powerful process because of the interdependence of information.

For example, atmospheric temperature, relative humidity, and density are related to one another. Consequently, assembling independent data sets of temperature and relative humidity allows one to assess the density without necessarily measuring it directly. As another example, the attitude of an individual responding to a survey on insurance costs may depend upon a recent experience, such as a hailstorm that destroyed the roof of their house. By synthesizing weather data and public survey responses, researchers can better understand what determines personal views and how those views change with time.

Most disciplines, especially in the social, behavioral, biological, and chemical sciences, as well as the humanities, offer entire courses in research methods (chapter 4). They focus on specific tools and approaches for gathering and analyzing data, particularly statistics; methods for constructing and testing hypotheses; underlying factors explaining why data are correlated; and the dangers of making false inferences. Indeed, many disciplines rely upon statistics, and their value lies in the fact that they provide a quantitative basis for estimating whether something likely happened for a reason, usually stated in a hypothesis, or rather likely occurred by chance.

Some interesting examples exist that show amazing statistical correlation without causality (e.g., Woollaston 2014). For example, how the sale of potato chips relates to people dying by falling out of a wheelchair, or how the sale of eggs correlates with the number of people killed in transport accidents. In both cases, no rational cause and effect exists whatsoever between the two quantities being compared. Statistics are incredibly powerful, but they sometimes are used ineffectively or inappropriately by well-intentioned researchers. If, as a detective, you use evidence incorrectly, you could get an innocent person convicted and lose your bar license. As a researcher, if you use statistics inappropriately, you could lose credibility, be subject to charges of research misconduct, or produce results that one day could lead to serious harm or even loss of life.

Statistics also can be intentionally misused to produce favorable results (section 4.7 and chapter 9), which is why a solid understanding of statistics is so important. Advanced tools such as data mining are exceptionally helpful for identifying patterns in data; but once again, causality must be established and the results should be explainable so as to ensure no bias or other problematic issues have arisen.

Another powerful analysis tool is computer visualization. Thirty years ago, visualization tended to be a novelty, used mostly to explain a complex phenomenon to a nonexpert. Today, visualization is used in every field, from classics and architecture to medicine and music. Powerful visualization tools allow researchers to bring many data sets together to understand their complex interrelationships visually, but also quantitatively, through the use of three dimensions, animation, tactile response, stereoscopic viewing, and sound. Ultimately, making sense of the information or data you have gathered, via the testing of your hypothesis through experimentation, is the end game of research. You are incredibly fortunate that more tools and information are available today to achieve this end than ever before.

Assess Your Comprehension

1. Why is placing your planned research in the context of previous work important?
2. What sources of information exist to help frame your research against previous work?
3. What is a primary source of information and how does it differ from a secondary source?
4. Why should you physically protect sources of information for your research and creative activity?
5. In what ways can you physically protect sources of information for your research and creative activity?
6. What is source validation and why is it important?
7. Provide a definition for data.
8. What are data QC and data QA, and how do they differ?
9. How can tools, such as artificial intelligence and machine learning, improve our ability to understand and synthesize data?

Exercises to Deepen Your Understanding

Exercise 1: Select a research topic of interest and perform a literature review or other appropriate analysis of work conducted previously (e.g., video and audio recordings of a performance, imagery if the topic involves art or manuscripts). Summarize your findings and, in so doing, identify important gaps in knowledge and understanding, and limitations of or flaws in previous work, as a means for suggesting future work. Additionally, create several questions you would seek to answer were you to prepare a research proposal on the chosen topic, and from them formulate a hypothesis. Although you may use informal sources, such as Wikipedia, to gain an overall understanding of the topic, your analysis should utilize mostly scholarly, archive sources of sufficient number to make thoughtful, thorough, and persuasive arguments.

Exercise 2: Artificial intelligence/machine learning is dramatically changing the landscape by which researchers can review and synthesize the scholarly record. Exercise 1 involves using the traditional "manual" approach of searching the scholarly record to understand previous work, identify gaps, and pose questions. For the present exercise, explore options for applying artificial intelligence/machine learning to achieve the same ends (you are

not asked to apply artificial intelligence, but rather explore capabilities and options). What services exist now, and in what disciplines are they available? Note that artificial intelligence does far more than identify articles relevant to your work, and in fact can synthesize findings across hundreds to thousands of articles, identify knowledge gaps, and even generate hypotheses. Summarize findings from your investigation and also describe capabilities that are on the horizon. How can artificial intelligence impact your own research?

Exercise 3: Data quality control (QC), data quality assurance (QA), and management are critically important in research and must be performed with great care. At https://drive.google.com/file/d/1_E6Wfg1nCI1UtAFIf6hqi8A2Uwpv492P/view?usp=sharing, you will find a spreadsheet containing time series data collected from a ground-based weather observing station during the passage of a tornado in Oklahoma. All quantities are labeled, and note that the wind direction is the compass direction from which the wind is blowing—such that a northerly wind is from 360 degrees, an easterly wind is from 90 degrees, a southerly wind is from 180 degrees, and a westerly wind is from 270 degrees. These are "raw" data from the sensors and have not been subjected to quality control processes. Examine the data to see if you can spot anomalies. For example, all values should be positive numbers, and dramatic changes from one time period to the next likely would reflect an error (for example, if the temperature dropped 20 degrees in five seconds with no substantial changes in other quantities at the same time). Also, an observation should be present for each time period shown or else be tagged as "missing data."

If you were developing an automated quality control algorithm, how might you go about identifying and correcting errors, including missing values? What factors might be responsible for creating errors or anomalies in the data? If you are interested in how data quality control and quality assurance are actually applied by the organization supplying the data, visit https://journals.ametsoc.org/view/journals/atot/27/10/2010jtecha1433_1.xml.

Exercise 4: Select a research problem of interest, perhaps from a previous exercise, and describe how you would go about creating (not collecting or finding) data to study it (in this context, data refers to information, physical artifacts, or other resources that support your work). Would you need to build an instrument, conduct interviews, make recordings, and take photographs? How would you go about applying quality control to the data to ensure accuracy and representativeness, and how would you make the data available to other researchers having similar interests? Be specific in your

answers and include information about locations or sources where such data could in fact be gathered.

Exercise 5: Describe the tools and methods used in your own research to synthesize data and develop understanding from them. Do you generate the data yourself, depend upon data from other sources, or use a combination of these approaches? What type(s) of data do you utilize (e.g., numerical, textual, graphical, animation, time series)? To what extent are the original data modified in the synthesis process, and what rules or procedures do you employ? If your synthesis/analysis involves multiple steps that produce modified data, which data do you save, and why? If you plan to make your data available to others, say to build upon your research, in what ways will you do so? Do mechanisms exist to archive your data in national repositories?

If you are not yet engaged in research, interview senior researchers and peers in your institution who have collected or are in the process of collecting data, ideally across multiple disciplines, and compare and contrast the approaches used.

The MIT Press would like to thank the anonymous peer reviewers who provided comments on drafts of this book. The generous work of academic experts is essential for establishing the authority and quality of our publications. We acknowledge with gratitude the contributions of these otherwise uncredited readers.

This book was set in Times Roman by Westchester Publishing Services, Danbury, CT.

Library of Congress Cataloging-in-Publication Data

Names: Droegemeier, Kelvin, 1958– author.
Title: Demystifying the academic research enterprise : becoming a successful scholar in a complex and competitive environment / Kelvin K. Droegemeier.
Description: Cambridge, Massachusetts : The MIT Press, [2023] | Includes bibliographical references and index.
Identifiers: LCCN 2022062131 (print) | LCCN 2022062132 (ebook) | ISBN 9780262547079 (paperback) | ISBN 9780262377218 (epub) | ISBN 9780262377201 (pdf)
Subjects: LCSH: Universities and colleges—Research—United States. | Research—Moral and ethical aspects—United States. | Research—Methodology. | Learning and scholarship—United States.
Classification: LCC LB2326.3 .D76 2023 (print) | LCC LB2326.3 (ebook) | DDC 378.0072—dc23/eng/20230512
LC record available at https://lccn.loc.gov/2022062131
LC ebook record available at https://lccn.loc.gov/2022062132