Foreword

The EDUCAUSE Center for Applied Research (ECAR) was launched on January 1, 2002, to create a body of research and analysis on important issues at the intersection of higher education and information technology (IT). ECAR is fulfilling its mission through a program of symposia and through the publication of biweekly research bulletins, detailed research studies, occasional papers, executive roadmaps, and case studies. These publications are designed to highlight effective practices, lessons learned, and other insights from the practical experience of campus leaders. Since ECAR’s inception, 14 symposia have been held, and more than 400 research publications have been issued.

The Challenges of Digital Data

The area of data management is particularly dear to me. Few people may know that unlike many, I began my career in IT management through the information resources management career pathway. Although most baby boomers in our community began their careers in programming, computer operations, production control, and other aspects of data processing, and although increasing numbers of today’s IT leaders are coming from the professoriate or other core constituencies of higher education, some of us come from the library, from records management, from archives administration, and from other cognate fields that deal more with the management and use of information than of technologies per se. In my case, I worked first as an archivist and then a records manager and was professionally preoccupied with the life cycle of information, regulatory compliance, information flows, information asset management, and the preservation of a meaningful and accurate historical record.

The introduction of computers and, later, networks has made information management at once harder and easier. Paper-based information artifacts take up a lot of space, but they can be relatively easily “marked” and “parked” for future retrieval. Further, our inability to control the management and use of individual words on individual pages made it possible for us to socialize administrative expectations that were in truth blunt. You were to either keep or discard all correspondence written between one date and another, pursuant to an institutional policy, or a local, state, or federal regulation. Records managers in the late 20th century could not find proverbial needles, so we contented ourselves with the management of haystacks instead.

As information became digital, the very basis of information management became
unclear. Documents have a clear structure, and their management can be and is defined and prescribed by that structure. Dimensions of document structure include date of origin, author, recipient, document type, length, and so forth. Information in digital form can be structured or unstructured. And even when it is structured, the dimensions that define this structure can be problematic. What, for example, is the unit of data management for the purposes of retention, destruction, storage, labeling, and so forth? Is it the data element? If we alter data elements, do we inadvertently obscure or even obliterate the meaning of other data elements that together constitute a greater information whole?

The digitization of information is dramatically changing the landscape of its management. Information that is digital is easy and relatively cheap to replicate. What is the record copy of information? The easy replication, transport, and storage of data have conspired to make information superabundant. What does superabundance of information mean in the context of laws, regulations, and practices that seek to create an accurate and coherent record of human activity? How many times have we seen the proverbial digital smoking gun disappear from the record only to be resurrected from an audit log, a backup file, or from files completely outside a coherent record system?

As we have undertaken to increase the human capacity to create, move, and store information, so have the media of storage available to us become more and more transitory. Cuneiform impressions on clay, carvings on cave walls, hieroglyphs on papyri, or inks on animal skins have endured for centuries or even millennia, whereas we struggle to read the digital record of the NASA Viking missions of 1976.

Digital techniques have lowered the cost of retaining information, giving rise to new economics of information. In the paper world, today’s news was tomorrow’s fish wrap. All but the most important historical documents—including most books—marched through a life cycle that imbued them with great value at their birth and then eroded that value quickly over time as owners and publishers moved to liberate precious real estate from the demands of “inventory.” In the digital context, multiple copies of data are regularly created for any number of reasons (for example, caching to accelerate search). As a result, whole businesses are arising to understand and exploit the residual value of information in the so-called long tail.

Digital information is not only relatively more ephemeral, it can also be altered by a variety of people. Ink on paper can be altered as well, but not easily and rarely without a trace. The ephemeral nature of digital information contributes to making its creation, management, and use a “team sport.” Archivists, in particular, and historians have long struggled with the concept of provenance—the lineage of a document. In the digital context, data can be mashed up and easily altered and can appear in a multiplicity of contexts. Understanding provenance, and therefore the accuracy, reliability, authenticity, and credibility, of information sources is made more and more complex and problematic. Not surprisingly, this problem is especially vexing in academic contexts. Do our historical concepts of sole authorship, originality, or plagiarism hold in this digital context?

The challenges of digital data go well beyond these few areas and examples. The complexity associated with managing data is increasing faster than our capacity to keep up.

This ECAR study chronicles the current state of the race between the capabilities that we now enjoy to create, manipulate, store, disseminate, and find information and our ability to organize, contextualize, interpret, or preserve that information.
Much Ground to Cover and Much Variation in Coverage

ECAR principal investigator Ron Yanosky has looked closely at five aspects of data management:

- data integrity and quality,
- analytics,
- data stewardship and security,
- content and records management, and
- management of research data.

Not surprisingly, Yanosky found tremendous variation in survey responses across these aspects of the problem. As the reader would expect, survey respondents from research-intensive universities expressed far greater concerns about an impending data deluge than did counterparts from associate’s or bachelor’s institutions. Notwithstanding many of the findings that corroborate our expectations, several stand out:

- The good news is that respondents are generally in agreement that their institutions have today and will continue to have in three years the IT infrastructure they need to manage data in administrative systems, learning management systems, web content, e-mail, research systems, video and audio files, and digital library systems. This confidence answers the question, Can you create it, move it through networks, compute it, find it, and store it?
- The less good news is that despite the progress that has been made in the past 20 years to bring the campus CIO to the cabinet level, managing the “I” in CIO continues to be a challenge. Respondents to this survey were not confident that the data in their systems are accurate, well defined, or synchronized properly across systems, nor were they confident that the processes to ensure things such as data quality are in place. Among our respondents the formal assignment of responsibility for managing data resources to data stewards was “the exception rather than the rule.”
- In administrative contexts we are still not a culture of evidence. On balance, our research participants described environments that are data rich but information poor. Poor data quality, a paucity of analytical tools, and a workforce that is largely unschooled in quantitative analytical methods conspire to yield a picture of relatively low adoption of analytics in higher education. Further, in cases where responses could be compared with those made to identical questions asked in 2005, we can only conclude that little if any progress has been made in this area of activity. These findings are in sharp relief to the vocal call among state and federal lawmakers and regulators to use data in more sophisticated ways to support decision making, student success, and other purposes.
- Not surprisingly, IT security—an area of rather significant investment over the past five years—is one of the areas that survey respondents were most satisfied with.
- Research university respondents were nearly unanimous in their expectations that research data volumes will grow significantly in the next three years. Keeping up with a possible deluge appears to be considered an economic issue more than anything.
- Survey respondents do not believe that their institutions are getting maximum academic and business value from the information they have. In fact, nearly half of the survey respondents disagreed or strongly disagreed with the statement “We get maximum academic (business) value from institutional data.” Although the standard
set in this question is a high one, it is clear from other responses that the situation regarding “value capture” is unsatisfactory.

**What Can We Conclude from This Report?**

We may certainly conclude that many of our IT leaders are charged as—or otherwise behave as—chief technology officers rather than chief information officers. The state of data management in this sample of higher education institutions is mixed. Most respondents appeared to be reasonably strong where it comes to managing an infrastructure for data and to securing that data. By the same token, many respondents were clear that data quality is lackluster, that little effort is made to mine institutional data to promote better institutional outcomes, and that although there is a lot of data to manage, the institutions are not getting full value currently from the data they have.

These findings leave me and likely many readers confused. Is this a scandal? A long-standing expectation? A critical weakness? Is higher education substantially different in this arena of endeavor than corporate organizations, governments, or others? Should these findings be the basis of a clarion call to action, or are they a reflection of a broader and ultimately sensible set of priorities?

My instinct on reading this fine report is that effective data management depends on ascertaining the dependencies among the aspects studied here. At bedrock, if regulators, educational leaders, politicians, and others want to use information to promote student success, to lower rates of student attrition, to improve institutional decisions, and to promote transparency, then attention must be paid first and foremost to data quality. Are the data in college and university systems untrustworthy, and if so, why? Building a superstructure of complex data mining, analytics, simulation, and benchmarking capabilities is likely wasteful at best if the underlying data cannot be trusted. Creating meaningful data quality is an enterprise activity that is reasonably well understood. It is also an activity that is rife with politics. EDUCAUSE clearly has an opportunity here to transmit best practices in the areas of data governance, knowledge bases, data architecture, quality assurance and management, and so forth. Campus IT leaders will need to build social networks that will make it possible to rationalize institutional dictionaries and to re-architect accidental management environments that contribute to low data quality. Yanosky provides helpful and practical summaries—with practical advice in every chapter.

**Too Many People to Thank**

The ECAR study of data management was an ambitious undertaking, and of course there are too many people to thank. Yanosky is a master at making sense of very complex data and at keeping his eyes fixed on big and actionable questions, even while wrestling with large volumes of cacophonous data. ECAR Fellows Susan Foster, Judith A. Pirani Pirani, Gail Salaway, and Mark Sheehan conducted qualitative research to gain deeper insights into the research data and to capture additional ideas and viewpoints for the study.

To enrich this core study, we are able to add two case studies. The first, *Revitalizing Data Stewardship through Risk Reduction: Managing Sensitive Data at the University of Virginia*, written by Bob Albrecht and Judith A. Pirani, presents the institution’s multilayered strategy that remediates past data usage practices, creates an updated data policy framework, and better protects university data by strengthening the IT infrastructure security. Pirani and Donald Z. Spicer delve into technical and archival issues surrounding the long-term digital preservation of research data in the second case study, about the
Chronopolis Digital Preservation Initiative, a multi-institutional grid-based project involving the San Diego Supercomputer Center, the University of California San Diego Libraries, the National Center for Atmospheric Research, and the University of Maryland’s Institute for Advanced Computer Studies. Many people graciously took time out of their busy schedules to assist us with these case studies, most notably James Hilton, Vice President and CIO, and Shirley Payne, Assistant Vice President for Information Security, Policy, and Records, at the University of Virginia; Brian Schottlaender, University Librarian, and Ardys Kozbial, Technology Outreach Librarian, University of California San Diego Libraries; and David Minor, Head of Curation Services, San Diego Supercomputer Center.

We are especially grateful to our colleagues at a number of institutions: Shelton Waggener, Associate Vice President and CIO at the University of California at Berkeley, and Bruce Maas, CIO at the University of Wisconsin–Milwaukee, provided much-appreciated advice about institutional data management strategy and practices. James Phelps, Senior Information Technology Architect at the University of Wisconsin–Madison, helped us understand the role of the data architect and current practices in that area, and Shirley Payne generously contributed to our understanding of data stewardship and security policy. Our ECAR colleagues Donald Spicer and Susan Foster, experienced CIOs both, reviewed our survey questions and made suggestions that greatly improved the study. Finally, Julie Curtis and her colleagues at SunGard Higher Education provided valuable insights about the technology and cultural issues affecting academic analytics and data management.

Of course, after all of this work on content development is complete, the work of the production team begins. The care of our investigators and fellows in constructing and designing surveys and in analyzing responses and checking analyses is matched by a team of editors under the guidance of Gregory Dobbin and Nancy Hays. They are thorough people and work with a team of editors, proofreaders, digital compositors, and printers. In studies where a quarter-inch shift in a column can obliterate a careful analysis, one cannot understate the effort these people make or the successes they claim. And last, but of course not least, Toby Sitko resides at the interface of the research team and the production team and orchestrates the overall project with the skill of a symphony conductor. ECAR depends on her every day.

Richard N. Katz
Boulder, Colorado