1 Executive Summary

Not for nothing is the modern era frequently called the Information Age. Computing and storage hardware, following Moore's law for more than half a century, has brought us the 32-gigabyte pocket device and the multiterabyte personal computer. We don't seem to have any difficulty at all finding new ways to fill these ever-fattening data piggy banks. The IT research firm IDC, taking a shot at the admittedly elusive question of how big the "digital universe" is, suggests a suitably astronomical size: 281 exabytes (billions of gigabytes) of data created, captured, and replicated in 2007 alone, growing at 60% annually.[1]

Higher education has played a big role in making this data "big bang" possible, and the purpose of this study is to examine its effects on colleges and universities and the ways their IT organizations deal with it. The challenges facing institutions may be understood by considering three broad areas of data impact.

First, like all big organizations running a complicated business, higher education institutions consume and generate a lot of operational information. Digital use of this kind of information goes back to the days when the 80-column punch card symbolized "data processing," and many of that era's classic problems remain with us today: making sure data is accurately recorded, kept up to date and in sync with business processes, and available for use beyond the narrow operational need it was originally created to serve. For the most part, this kind of data is *structured*, i.e., modeled according to a rigorous scheme that defines its size, its type, and its legitimate values. Though modern relational databases and enterprise applications suites have made operational use of business information much speedier and more flexible, they have not been as successful at opening it up for broader analysis and management use. It remains much easier today to place the right student in the right course than to use aggregated information about students and courses to do enrollment and curriculum planning. Yet funders and accreditors increasingly ask for metrics that permit commensurate comparisons of unit and institutional performance, and higher education reformers are envisioning a world of outcomes-based assessment and easily exchanged student records. To the extent that these trends continue, the quality of operational data and the ability to use data to analyze and manage will become ever more significant.

A second domain of data challenge is the enormous body of data known as *content*. Fed by the intersection of new networked modes of delivering information (above all, the web) and by the shift of all kinds of media

toward digital formats, content is a necessarily vague category yet an essential one to manage effectively. At a modern university, content includes not only the material available on the institution's website(s) but also documents such as memoranda and spreadsheets, its publications, the institution's online teaching materials, a growing proportion of its library and museum collections, and much else besides. Most content is unstructured, meaning that it contains more or less free-form information rather than the sort of modeled data consumed in a typical business application. Because content is often formatted for use in website templates or for recall from a database, however, it increasingly takes the form of semistructured data. Much content—a policy statement, a website describing academic programs, a course catalog entry—requires some kind of official approval; much of it also needs to be formally described and cataloged in order to be trackable and searchable and available when needed. And much of it is ephemeral junk. The great challenge of institutional content management is to provide the tools that allow the right people to create, publish, find, and preserve or winnow the right content according to the needs of the institution. Yet the technology for handling unstructured or semistructured data remains far less mature than that for handling structured data.

A final domain of the higher education data challenge is research data: Whether from experiment or computation, modern science is producing enormous quantities of it. Digital data is more portable and searchable than traditional forms of data capture—so much so that some science visionaries, backed by funding agencies, believe that the time has come for a new era of shared data cyberinfrastructure and open data access. Yet besides sharing many of the problems of operational business data and content, research data raises unique problems relating to ownership, preservation, and interpretation. Solving these problems could enable a dramatic increase in research productivity; failing to solve them could mean that much of the data now being collected with such creativity and effort could be lost through deterioration or simply because nobody knows it exists.

Our capacity to produce data and expose it over networks has far outstripped our ability to reform or replace the business, legal, and cultural practices that defined our relationship to information in the era before data superabundance. As a result, requirements relating to new forms of data are piled onto ongoing challenges from more familiar forms, while government and institutional leadership subject data to new kinds of regulation, often in a reactive way not very well informed by an appreciation of data management principles.

Our study looks at how higher education institutions are responding to these challenges, with particular focus on institutional content, research data, and operational business/academic data and its analysis. Our aim has been to provide institutional CIOs and other concerned parties with information about the state of higher education data management and to identify practices associated with good data management outcomes.

Defining Institutional Data Management

In the context of this study, *institutional data management* refers to the policies and practices by which higher education institutions effectively collect, protect, and use digital information assets to meet academic and business needs. It's worth noting that though our overall guiding concept is data management, our definition refers broadly to information assets in digital form. Although some authors make a point of distinguishing between data (symbolic elementary units of information) and information (data that is interpreted, processed, or put to a purpose), we maintain that a study focused on the

practical concerns of institutional CIOs will lose more than it gains by insisting on too sharp a distinction between the two. We believe that the term *institutional data management* is elastic enough to reflect our interest in enterprise information needs while speaking the language of working IT professionals.

Methodology

Our study of institutional data management in higher education took a multipart approach that consisted of

- a literature review to examine data management frameworks and definitions, to define issues and establish research questions;
- consultation with higher education IT administrators and data management experts to identify and validate survey questions;
- a quantitative web-based survey of EDUCAUSE member institutions that received 309 responses, 78% of which were from the institutional CIO or equivalent;
- qualitative interviews with 23 higher education IT leaders and staff; and
- two case studies, one examining a sensitive-data risk reduction initiative at the University of Virginia, and one concerning Chronopolis, a multi-institutional grid-based project for the long-term preservation of research data.

Certain questions in our quantitative survey were identical to questions posed in several earlier ECAR surveys. Where applicable, we carried out a longitudinal analysis of responses to these questions on the basis of institutions that had responded to both the earlier and the current survey.

Key Findings

Our study was chiefly concerned with how institutional IT units are coping with an explosion in the volume and variety of data that they have to manage, and how well they think they're doing at such data management goals as keeping data secure and getting maximum academic and business value from it. We drew eclectically (and selectively) from a number of data management frameworks and standards to identify topics of research interest. Areas of survey coverage, which map roughly to the organization of this study report, included the following:

- *Institutional context*—size and scale of the institutional data environment, data-related staffing, institutional drivers, and barriers to data management.
- *Data integrity and quality*—planning and control activities that measure and ensure the fitness of data for use, respondent assessment of institutional data quality, and management of "master" data.
- *Analytics*—processes to provide decision support (aka business intelligence) for institutional leaders and other campus management and staff; and supporting infrastructure, including data warehouses, data marts, and related tools.
- *Data stewardship and security*—assignment of responsibilities for the security and quality of data; exercise of authority and decision making over the management of data assets and creation of data policy; and related data policies.
- *Content and records management*—infrastructure and practices to store, access, and (as applicable) archive data within content systems (e.g., web, e-learning, documents) and institutional records systems; and assessment of the institution's compliance with records retention requirements.
- *Research data management*—distribution of research-related data responsibilities among institutional units, research data ownership and other

research-related policies, and institutional ability to support the long-term preservation of research data.

- *Data management outcomes*—respondent assessment of institutional performance on summary measures of data management effectiveness.

In the following sections, we summarize and synthesize our main findings.

Drowning in Data?

Clichés about today's data environment often invoke catastrophic images involving water. But are higher education institutions "drowning" in data, as is often said? Our respondents certainly reported high levels of data growth. When we asked them to characterize changes in the volume of institutional data in the past 12 months using a scale from 1 (greatly decreased) to 5 (greatly increased), the mean response for seven out of eight categories of data was at or slightly above 4 (increased). The fastest-growing types of data by volume were learning management systems data and content (4.22) and e-mail (4.18), and the slowest-growing was research data (3.74). The apparently low figure for research, however, was chiefly an effect of combining research-oriented institutions (4.27) and teaching-oriented ones (3.52).

Respondents were fairly positive in assessing whether their institutions have the infrastructure needed to effectively manage various kinds of data (see Figure 1-1). Except for research data and video/audio content available online—two types of data that are particularly hungry consumers of data storage and bandwidth—respondents tended toward agreement with this proposition, and they were still more positive about their infrastructure in three years. About half of the central IT units among our respondent institutions managed a total disk storage capacity of 20 terabytes or less, and the median estimated growth in that

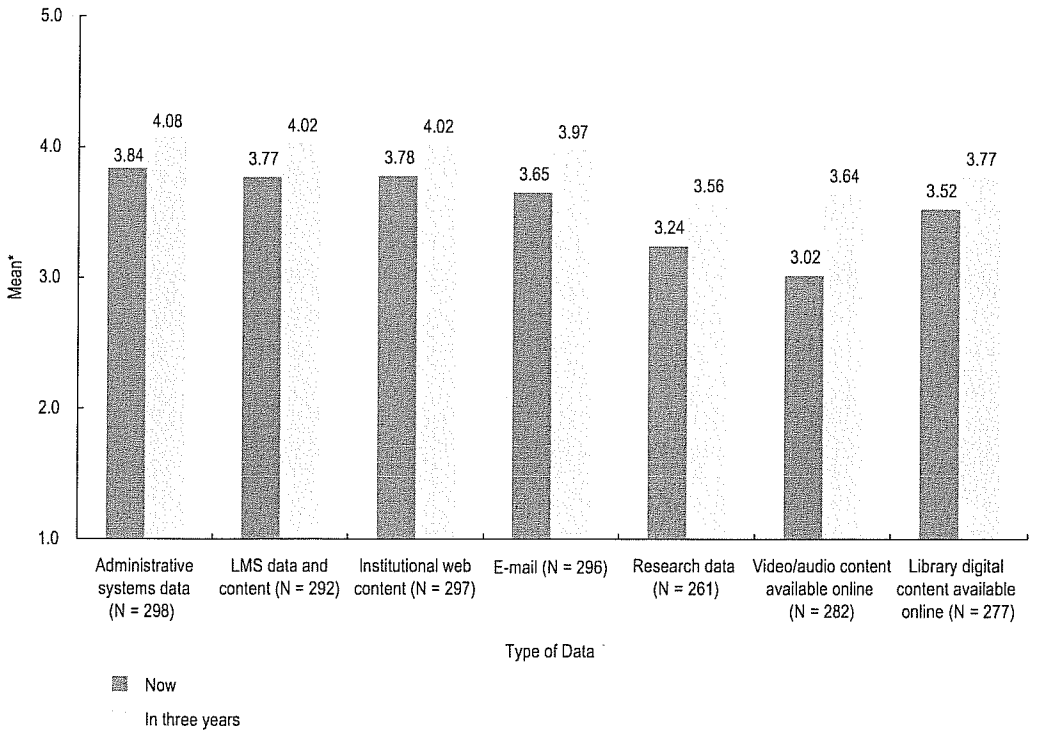

Figure 1-1. Institution Has Infrastructure Needed to Manage Institutional Data, Now and in Three Years

**Scale: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree*

capacity over the next 12 months was 20%. In short, though our respondents report data volume growth and plenty of it, they don't seem to indicate that they're "drowning," at least in terms of capacity and infrastructure.

So where are the pain points? Lack of funds was by far the barrier to investment in institutional data management that our respondents most often named, followed by lack of staff expertise and a decentralized or informal institutional culture. As for investment drivers, institutional "sticks" (improving data security and regulatory compliance) stood somewhat ahead of "carrots" (improving business/academic unit operational efficiency and enhancing data-driven decision making).

Identity management is a crucial enabler of good data management because so much about data security and appropriate use hinges on the ability to authenticate and authorize users. Our respondents, however, averaged only an uninspiring "neutral" response when we asked whether they agreed that their institutions' identity management provided sufficiently granular data access authorization.

Enterprise Data Quality

What makes data useful? Many things, no doubt, but common to most business uses of data would be its accuracy, along with confidence that it is well defined and stored in the right place, that its value is in sync everywhere it's used, and that it resides in an environment attuned to good data quality.

One of the major themes of our study is that these questions of data quality seem to be foundational to many aspects of data management. And as Table 1-1 shows, when we asked respondents a series of data-quality-related questions pertaining to major data elements within their institutions' major administrative systems and data stores (such as HR, finance, and student administration

Table 1-1. Administrative Enterprise System Data Quality Measures

	Mean*	N	Std. Deviation
Characteristics of Major Data Elements			
A system of record is identified for each major data element.	3.54	301	0.981
Each major data element has a single definition that is recognized across the institution.	3.29	308	1.117
Each major data element is coded consistently across systems and data stores.	3.12	301	1.075
When the value of a major data element changes, the change propagates across all enterprise systems and data stores that use it.	3.11	308	1.140
When the value of a major data element changes, the change propagates across all business/academic unit ("shadow") systems and data stores that use it.	2.52	299	1.082
Data Quality Processes			
Processes are in place to assure data quality at the point of capture/origin (e.g., data entry).	3.24	306	0.975
Automated systems are in place to validate data across enterprise systems and data stores.	2.89	302	1.035
Processes are in place for documenting and reviewing all identified data quality issues.	2.76	304	0.995

**Scale: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree*

systems), they were far from enthusiastic. Mean responses on a scale from 1 (strongly disagree) to 5 (strongly agree) were for the most part only slightly above neutral, and in some cases were below neutral.

We averaged each institution's responses to these questions to create a single enterprise data quality score. This score, which for all respondents averaged a middling 3 on our 5-point scale, proved to be positively associated with many other measures in the study, suggesting (though not proving) that better enterprise data quality is a factor in better data management outcomes. We also found some evidence that perceived enterprise data quality is susceptible to educational efforts: Where respondents agreed more strongly that executives, managers, and staff receive training in data quality issues, enterprise data quality scores tended to be higher.

Data Analytics

Modern administrative systems promise not just to handle real-time transactions such as paying a bill or enrolling a student in a course, but also to make the aggregate information collected by these systems available for institutional management and planning. How institutions approach reporting, modeling, analysis, and decision support—or more simply, analytics—was one of the key concerns of our study.

Effective use of analytics is often said to be as much a cultural matter as a technological one, because institutional executives and managers simply may not be accustomed to using empirical data to make decisions, or they may wish to restrict access to relevant data. Our respondents, however, were fairly positive about such cultural factors. Majorities of them agreed or strongly agreed with such propositions as that their institution's leadership was committed to evidence-based decision making and that their institution was committed to making management information widely available.

At the same time, we found a fairly modest infrastructure for analytics. Of four different kinds of analytics-related data stores we asked about, including data warehouses, data marts, and operational data stores in each of two different usage modes, 41% said they had none of them in use, and another 18% reported only one. Roughly one-third reported data warehouses in use institution-wide, with a few more reporting them in use at the school, college, or department level. The incidence of most of the analytics infrastructure items we asked about was unchanged since ECAR's 2005 survey on academic analytics.

As in that study, we found that the most common primary use of analytics tools by far was for extraction and reporting of transaction-level data. Advanced uses of analytics such as identifying at-risk students or auto-triggering business processes were widespread in terms of the percentage of institutions using them at least occasionally, but relatively few said they made use of them "usually" or "almost always." Higher-frequency use of advanced analytics, however, proved to be positively associated with a number of good data management outcomes, as we report later in this chapter.

Data Stewardship and Security Policies

Authorities on data management stress that data should be "owned" by the people who know it best and have the best incentives to care for it and keep it secure. Our results relating to the division of data responsibilities suggest that this advice is being taken seriously at most of our respondent institutions. When we asked whether primary responsibility for various data-related items lay with business/academic units, central IT, or neither, we found a generally reasonable division of labor. Business and academic units most often had primary responsibility for data accuracy, fitness for purpose, and data access decisions, and central IT most often

was responsible for specifying data formats and metadata management. One surprise was that at two-thirds of institutions, central IT was responsible for notification of affected parties following a data breach—perhaps an endorsement of the virtues of having a single central party responsible for this delicate task, rather than a multitude of business data owners.

Given this rather level-headed distribution of tasks, we were somewhat surprised to find that another frequently mentioned best practice—formally assigning responsibility for managing data resources to data stewards—was the exception rather than the rule. Overall, only about a third of institutions reported having a documented institutional policy defining data steward responsibilities, though slightly more than half of institutions with FTE enrollments greater than 15,000 had them.

Institutions with data stewardship policies did a little better than those without them in enterprise data quality score, and they were much more likely to say they had a formal process for classifying institutional data by risk or sensitivity level (another commonly mentioned best practice). However, we were surprised to find few other meaningful associations with data stewardship policies; in particular, this was not a differentiating factor in the major data management outcomes we asked about.

Other sorts of policies relating to data use, security, and access were more common than those for data stewardship. About three-quarters of institutions said they have a documented policy on the acceptable use of institutional data, and about 6 in 10 said they have one defining individual employee responsibilities for data security practices. It may be that some institutions see these as alternatives to a formal data stewardship policy, though in both cases institutions reporting a data stewardship policy were also more likely to have one of these two policies.

Institutional Content Management

It's increasingly difficult to think of some aspect of institutional work that doesn't create a digital document, spawn an entry on a website, get communicated by electronic means, or somehow contribute to the body of digital objects known loosely as "content." Some portion of this digital content meets the business, regulatory, or historical significance criteria that make content into records. Our study looked at a number of facets of institutional content management, including the overall environment for handling content, management of and plans for the institutional website, and the institution's ability to meet the requirements of records management.

IT analysts often promote the notion of enterprise content management, that is, integrated systems and processes for dealing with a wide variety of digital content relevant to business operations. Only about 12% of our respondents, however, said that their institution has an integrated enterprise content management solution managing the widest possible range of content (see Figure 1-2). Most often, they said that they have a mix of enterprise and local best-of-breed solutions varying by content type (36%), and about 40% reported no enterprise solutions for content, only local/departmental solutions or a generally ad hoc approach to digital content management. Looking ahead three years, most institutions expected to move in a more enterprise-oriented direction, either with an integrated enterprise content management system (32%), an enterprise best-of-breed–predominant environment (28%), or a mixed enterprise/local best-of-breed environment (31%).

Web content management was a particular focus of our survey. Respondents averaged roughly neutral responses to a statement that the institutional website routinely exceeds the expectations of constituents; they were most

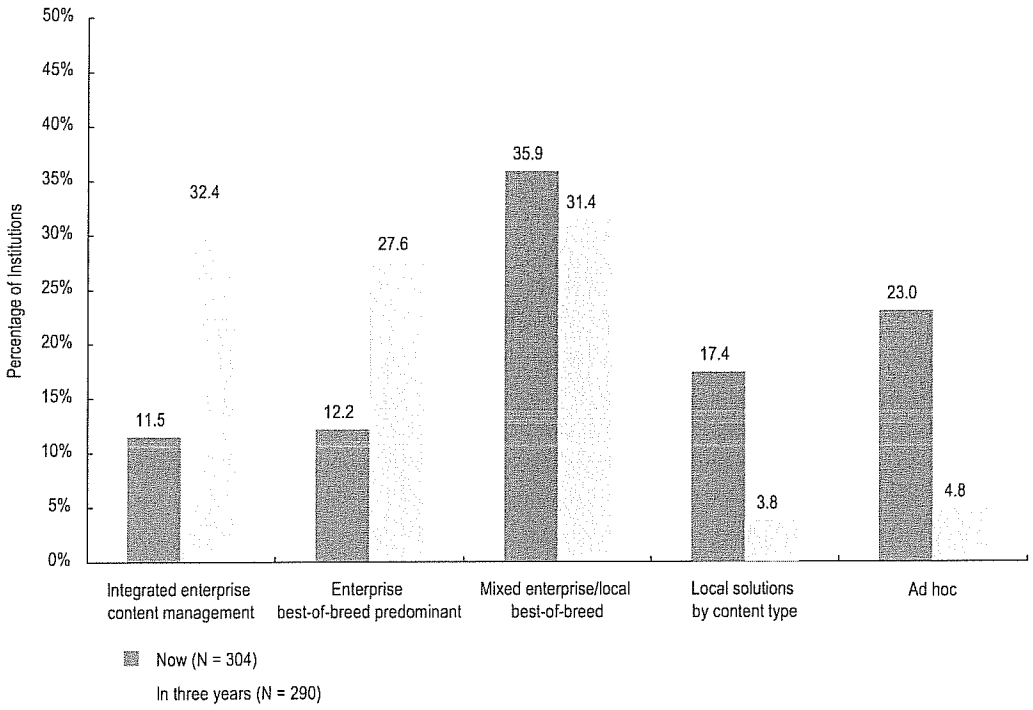

Figure 1-2. Institutional Content Management Environment, Now and in Three Years

positive with respect to public users (mean 3.23 on a scale of 1 to 5) and least positive about faculty users (mean 2.80). Perceptions of exceeding constituent expectations were associated with a number of web content management best practices that we asked about. For example, where respondents said that characteristics such as a database-driven web environment or consistent look and feel of web pages were more prevalent, they also tended to agree more strongly that their institutional website exceeded constituent expectations.

About half of institutions reported either an archivist or a records officer responsible for electronic records, a group responsible for overseeing electronic records management, or both. Respondents were not particularly positive about their institution's ability to comply with records retention and disposition requirements throughout the institution; those who characterized that ability as poor or fair (53%) slightly outnumbered those who rated it good, very good, or excellent. Institutions that reported e-records retention schedules in more areas of content, and those with higher enterprise data quality scores, tended to rate this ability higher, but somewhat surprisingly we found no difference based on the existence of an archivist/records officer or a group overseeing electronic records management.

Research Data Management

The capacity of digital technologies to capture research information in a form that can be readily analyzed and shared inspired the National Science Foundation to observe in 2005 that "digital data collections are at the heart" of "fundamentally new approaches to research and education."[2] They are also, however, at the heart of an emerging debate about how to store, share, and properly manage this huge, rapidly accumulating, and highly diverse body of data. Our study examined the context of digital research data management at respondent institutions, with a particular focus on the long-term preservation of data.

We began by asking about the direction of change in the basic IT infrastructure of research, including the use of high-performance computing and networking and the amount of research data storage. Overall, about 39% of institutions reported that use of high-performance computing had increased or greatly increased in the previous three years, and slight majorities reported such increases in high-performance networking and research data storage. Increases were much more common among research-oriented institutions than among teaching-oriented ones, and expectations of increased use in the next three years were more dramatic still: 98% of research-oriented institutions expected to increase or greatly increase amounts of research data storage over that time period.

When we asked respondents how often various campus entities were involved in research data support activities, we found widespread and rather diffuse participation in different activities. The activities central IT most often supported to a large or very large extent were data storage and data backup/recovery—classic central IT tasks—though about one in five institutions said that central IT provided assistance with creating research data management plans and support in the selection and use of research tools. The role of individual investigators, labs, or teams in providing their own research data management support was considerably higher at research-oriented institutions than at teaching-oriented ones.

Respondent institutions did not provide much evidence of aggressively undertaking long-term research data preservation activities, though our results were somewhat diluted by an unusually large proportion of "don't know" answers. Only about 31% of respondents said their institution assumes responsibility for archiving research data after investigator projects conclude, and most of these said the institution does so only in a few or some cases. A 41% "don't know" response, however, suggests that many respondents were not acquainted with their institution's activities in this area. When we asked whether respondents agreed that their institutions have the necessary commitment and funding mechanisms to support the long-term preservation of research data, their average answers fell well short of even a neutral (3) on our 5-point agreement scale. Two-thirds disagreed or strongly disagreed with the funding mechanisms item. Those institutions where central IT was more involved in providing research data management support tended to have somewhat more positive views.

Given these results, it's not surprising that respondents were rather lukewarm when asked if they agreed that their institutions met the data-related needs of institutional researchers. The average response overall was almost exactly neutral, but doctoral institutions averaged well below that; their mean agreement was 2.54, about midway between a disagree and a neutral response. As with the long-term preservation items mentioned above, agreement about meeting the data-related needs of researchers tended to be higher where central IT was more involved in research data support activities.

Data Management Outcomes

Data management is an especially complex area of IT practice. To tie its many facets together and determine how well respondents think their institutions are achieving goals that good data management ought to serve, we asked them to state their level of agreement with a series of statements about data management outcomes (see Table 1-2).

Given the amount of angst that has been expressed about IT security in the past decade, we were surprised to find that institutions agreed most strongly with the statement that restricted or sensitive data was secure at their institutions. This was the only outcome that

Table 1-2. Assessment of Data Management Outcomes

At my institution...	Mean*	N	Std. Deviation
Restricted/sensitive data is secure from unauthorized access.	3.87	305	0.770
We can support anticipated growth in the volume of data over the next three years.	3.32	304	1.053
Employees understand their responsibilities in the use of data.	3.14	304	0.898
We get maximum academic value from institutional data.	2.72	303	0.936
We get maximum business value from institutional data.	2.67	304	0.963

**Scale: 1 = strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree*

came close to a mean agree (4) response. Respondents were also relatively positive regarding their institutions' ability to support anticipated growth in the volume of data in the next three years—another item, given the "deluge" theme often associated with contemporary data management, that we might have expected to take a lower relative place in the outcomes.

By contrast, the two outcomes with the lowest average responses were the ones about getting maximum academic and business value from institutional data. Granted, these statements, by referring to "maximum" value, set a high standard, but it was striking that both fell well below even a neutral mean response. Nearly half of respondents disagreed or strongly disagreed with each of these statements.

Despite the diversity of concerns that our outcomes statements represented, we found a number of factors that were positively associated with most or all of them. Most interesting, and one of the critical findings of the study, was a strong relationship to enterprise data quality score: Institutions with higher scores tended to agree more strongly with all five of the outcomes items. These differences could be dramatic; for example, although agreement among institutions at the low end of our enterprise data quality score range averaged 2.01 on the maximum business value statement, those at the high end averaged 3.21.

Among the other factors we found positively associated with better outcomes measures were

- stronger agreement that the institution provides effective training to users of analytics tools (all outcomes except the data security item),
- an early adopter approach to technology adoption (as opposed to mainstream and late adopters), and
- stronger agreement that the institution has the necessary commitment to support long-term preservation of research data.

In addition, where institutions made frequent use of advanced analytics practices, and where they agreed more strongly that institutional leadership was committed to evidence-based decision making, they tended to report higher agreement about getting maximum academic and business value from institutional data.

Conclusion

As with a lot of higher education IT concerns, data management has the aspect of a workaday task being handled competently by busy IT departments with a lot of other tasks to perform. We didn't conclude that institutions are losing a desperate battle with a data deluge, though we did note that they seem to be more worried about handling the *variety* of data that the modern environment involves, as opposed

to its volume. And that makes sense: So far, at least, Moore's law has held steady for the density of storage technologies, whereas our mastery of data complexity—figuring out, for example, how to search a video file or cost-effectively tag thousands of documents with useful metadata—is evolving a lot more slowly.

In fact, it was not the poster concerns of data volume and security that produced the most notable results of our study, but the more mundane and operational ones. Institutions gave themselves relatively weak marks for data quality, and their analytics infrastructure and use seem not to have changed much since we investigated those topics in 2005. Yet our summary enterprise data quality score turned up again and again as a positive factor in desirable data management outcomes: Institutions with higher data quality scores gave better marks to their identity management capabilities, their institution's analytics training, their ability to manage all the varieties of data and digital content their institution needs, their websites' ability to exceed the expectations of users, and all of the general outcomes shown in Table 1-2. Likewise, effectiveness of analytics training proved to be one of the most powerful factors in data management outcomes, perhaps because it speaks of a commitment to *do* something with data, and use of advanced analytics techniques was strongly associated with two of the most important—and lowest-rated—outcomes, getting maximum academic and business value from institutional data.

That is, after all, what it's there for. In the Y2K era, higher education interest and investment in enterprise administrative systems surged, and at about the same time IT security and data privacy emerged as the great hot-button issues of IT administration. Each of these initiatives can claim its successes, even if they may be hard to recognize in the busyness and confusion of day-to-day IT operations. But it may be time to declare that success more explicitly and consider how institutions can work toward their goals by giving as much attention to data's long-term value as to its real-time transactional value, and by investing in the ability to analyze data as well as the ability to lock it down.

To extend attention from security and transactional systems to include data quality and analytics would, we believe, help institutions get more business value from their structured institutional data. As for the much more amorphous yet vital realm of content, it seems only fair to note that if institutional content management capabilities seem pretty immature, so too do content management technologies themselves. Our study, however, suggests some practical steps that institutions can take toward taming the content beast: approaching content from an enterprise view, being diligent about identifying and documenting the subset of documents that needs special treatment as records, and (to return to our favorite theme) maintaining a high-quality data environment.

As for research data, institutional and indeed scholarly needs are still inchoate, and though we noted some outlines of an institutional response to these needs, it's not at all clear that higher education institutions are committed to shouldering the burden of research data preservation, or even that they should. Just the same, we see some basis for IT attention to institutional research data needs, partly because our respondents on the whole seem to think their institutions are doing a mediocre job supporting their researchers, and partly because IT has skills—including data center administration, managing high-reliability applications, and large-scale data administration in an on-demand environment—that few other units on campus can boast of. At the very least, central IT can be an advocate for ensuring that the necessary expertise exists somewhere on campus, and IT can be a partner in contributing to its success.

Endnotes

1. John F. Gantz et al., *The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth through 2011* (Framingham, MA: IDC, 2008), http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf. This IDC white paper was sponsored by the storage products and services company EMC.
2. National Science Board, *Long-Lived Data Collections: Enabling Research and Education in the 21st Century* (Washington, DC: National Science Foundation, 2005): 9, http://www.nsf.gov/pubs/2005/nsb0540/.