Introduction and Methodology

A wealth of information creates a poverty of attention.
—Herbert Simon

In the second decade of the Internet era, the world is awash in a stunning quantity and variety of information. Exactly how much digital information we’re producing and living alongside is a tough question to answer, but there’s no doubt that the answer involves the kinds of numbers we used to associate mainly with extragalactic distances. A study by the University of California, Berkeley, estimated that in 2002 a total of 5 exabytes (5 billion gigabytes) of new data was recorded on media (including paper)—37,000 times the size of the Library of Congress—and another 18 exabytes flowed through electronic channels. The IT research firm IDC, estimating the size of what the company calls the digital universe, produced an even more eye-popping figure: 281 exabytes of digital information created, captured, and replicated in 2007 alone, a quantity IDC expects to grow by 60% annually through 2011.

These numbers reflect a particularly fast-growing component of image and video capture, but there seems to be hardly any endeavor that isn’t confronted with an exploding body of digital material. Scientific data captured in experiments or generated in computational research has reached heroic proportions, with no end in sight; at the upper end of data density, the Large Hadron Collider at CERN is expected to produce 15 million gigabytes of raw experimental data per year, a DVD every five seconds. The more bookish and text-bound side of the academy is undergoing a digital revolution of its own, as Google and other entities produce a searchable online corpus of books that could only have been found at the finest research libraries a decade ago—if there. Perhaps the most remarkable efflorescence of digital content has come from a combination of widening Internet access and web interactivity, a phenomenon almost impossible to quantify but suggested in outline by one estimate that almost 1.7 billion people worldwide now have access to the Internet, and another estimate that the number of available websites exceeds 225 million. And that doesn’t count more than 400,000 terabytes of new information generated annually by e-mail!

Nor are more traditional sorts of structured business data lying still. In a 2007 study by the IT advisory firm Ventana Research, more than half of organizations surveyed reported that their structured business data was growing by 20% or more annually, and their unstructured content was growing faster still. For that matter, such neat categories as structured and unstructured data are getting intertwined, as back-end databases dynamically feed websites and XML-related technologies contribute to...
what one analyst calls “a blurring distinction between data and documents.”

The sheer volume of information is only the beginning. As data flows with greater speed and ease, it overwhelms the physical and social restrictions that once kept it private or profitable, and those traditional constraints have increasingly been replaced by waves of privacy, security, and property rights regulation. What were once adequate tools for categorizing and storing information have been swamped, creating the need for new ways of describing and discovering information and of integrating its bewildering variety of expressions. As data gets more plentiful, ways of combining it grows exponentially, and people and institutions are asked to do more with it: to glean more from it; to put all of it at every (legitimate) user’s fingertips; and to link it, sync it, and use it to manage.

Higher education institutions aren’t the only organizations facing these challenges, but their historic mission of producing and aggregating information, along with the complexities of institutional management in an era of information overload, makes them especially sensitive to them. It seems that every aspect of IT administration, from the most traditional (major administrative applications) to the most emergent (online video and support for computational research), brings its associated bag of tricky data-related problems.

What’s more, with their often distributed organizational structures and highly expressive, data-loving constituents, higher education institutions can rarely impose centralized order on institutional data chaos even where that might seem like a good idea. The daily press of urgent business also seems to militate against attending to data issues that require a good deal of coordination and investigation even to describe, let alone solve. A survey conducted by the higher education IT architects’ organization ITANA in 2008 found that when asked to characterize the maturity of various aspects of their data management practice on a scale from 1 to 10, most institutions answered with a 5 or less in such areas as data quality and master data management, data warehousing and business intelligence, document/content/records management, and metadata management.

**Data Management Concepts and Definitions**

The body of IT expertise and practice surrounding these issues is variously known as data management, information management, enterprise information management, or some close variant. Some authors stress the conceptually useful distinction between data (symbolic elementary units of information) and information (data that is interpreted, processed, or put to a purpose). Attempts to carry the hierarchy further, to what systems theorist Russell Ackoff denoted as knowledge and wisdom, have been influential in management theory, though working “knowledge management” systems are not common in most IT environments, and wisdom management still appears to be a job for elders and philosophers, not IT professionals.

In practice, however, these distinctions may be more refined than necessary. It is clear that in an increasingly digital world, IT expertise surrounding data is the foundation for an environment from which individuals can either directly access information or create it through analysis or transformation. It’s interesting to note that DAMA International, a certification organization for data professionals, describes data resource management as “organizing people, energy, practices and procedures into activities that properly manage the information needs of an enterprise.” Our approach in this study has been to presume the fluid relationship between data and information, primarily employing a data-oriented vocabulary while still being mindful of the information needs of
Institutional Data Management

the institution and the reasons why IT leaders are called chief information officers.

As the guiding concept for this study, ECAR uses the term institutional data management, which we define as the policies and practices by which higher education institutions effectively collect, protect, and use digital information assets to meet academic and business needs. We believe that this term stretches well enough to reflect our interest in enterprise information needs while speaking the language of working IT professionals.

Study Scope and Objectives

In a world where all manner of processes and types of content are being virtualized, a study of data management runs the risk of being a study of everything. We therefore looked for frameworks that could suggest some structure and boundaries to the project.

Data management is not governed by a single controlled vocabulary or by well-established comprehensive frameworks or standards. The most thorough framework is DAMA International’s Data Management Body of Knowledge (DAMA DMBOK), which divides data management into 10 different functions. The DMBOK framework, however, pays little explicit attention to strategic or regulatory issues that are likely to be of concern to IT leaders, and it is still in draft form. Many implications for data management are distributed throughout the IT Governance Institute’s COBIT 4.1 framework, but its “manage data” control process is narrowly concerned with backup, storage, and disposal of data. The ISO/IEC 27002 information security standard (formerly ISO 17799) covers security issues relevant to data management, but no comprehensive ISO standard dedicated specifically to data management exists.

The absence of a single dominant model of data management has led us to adopt an eclectic but selective approach to the topic that draws both on the assorted models mentioned above (especially DMBOK) and on ECAR’s understanding of the particular concerns of higher education, informed in part by prior ECAR research. Our study focuses on the following main topics:

- 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 of activities that measure and ensure the fitness of data; 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 (for example, 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; processes for determining investigator compliance with research data management...
regulations; 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.

We have chosen not to include a number of topics that might be considered aspects of data management, either because they are things institutions have largely mastered or because ECAR has recently conducted studies related to them. These include database operations, backup/recovery, business continuity, and IT governance.

**Research Approach**

Our research consisted of four different components: a literature review, a quantitative web-based survey of IT leaders at EDUCAUSE member institutions, qualitative interviews with CIOs and others at selected institutions chosen on the basis of their responses to the survey, and case studies.

The literature review helped us determine workable definitions, and a reasonable scope, for this notoriously expansive concept, and it suggested hypotheses we could test while providing secondary studies for context and support. We examined articles and monographs from journalistic, academic, IT practitioner, and IT advisory sources. We also investigated relevant frameworks and standards, including DAMA International’s DAMA-DMBOK Functional Framework and the IT Governance Institute’s COBIT 4.1 framework.

With input from a number of higher education CIOs and data management experts, we designed a web-based survey aimed at the senior-most IT leader at respondent institutions. A copy of the survey resides at http://www.educause.edu/ECAR/InstitutionalDataManagementSur/163709. The survey was released in early February 2009 and remained in the field for approximately four weeks. Appendix A lists institutions that responded to the survey.

ECAR used qualitative interviews to gain insight into the quantitative results and to capture additional thoughts and ideas we would not have been able to explore through the survey alone. A total of 23 IT leaders and staff representing 19 institutions took part in our interviews; they are identified in Appendix B.

Finally, this study is accompanied by case studies that look in depth at data stewardship at the University of Virginia and at the Chronopolis project, a multi-institutional project sponsored by the Library of Congress that aims to create a grid-based environment for the long-term preservation of research data.

**Classification Schemes**

For comparison, we grouped institutions using categories derived from the 2000 edition of the Carnegie Classification of Institutions of Higher Education, developed by the Carnegie Foundation for the Advancement of Teaching. To obtain adequate numbers for statistical and descriptive purposes, we collapsed the Carnegie 2000 classifications as follows:

- **Doctoral institutions (DR).** The study grouped the doctoral-extensive and doctoral-intensive universities together.
- **Master’s institutions (MA).** The study grouped master’s colleges and universities I and II together.
- **Baccalaureate institutions (BA).** The study combined the three Carnegie 2000 baccalaureate groups together.
- **Associate’s institutions (AA).** These are the same as the Carnegie 2000 associate’s category.

To characterize respondent demographics, we also report an “other Carnegie” category that includes specialized institutions and U.S. higher education offices. Owing to the diversity and small size of this category, it does not
In our detailed data analysis by Carnegie class. We also tracked Canadian institutions in a separate, single category.

In November 2005, the Carnegie Foundation for the Advancement of Teaching introduced a new classification scheme employing additional institutional characteristics. We have not provided a crosswalk to the new scheme, in large part because we suspect that our readers will be more familiar with the older 2000 taxonomy.

We also follow the lead of several previous ECAR studies by analyzing results according to respondent institutions’ self-classification regarding the relationship of research and teaching in their institutional missions. As Table 2-1 shows, we asked respondents which of four statements best described their institution, using the terms research essential, balanced, teaching favored, and teaching essential as short names for each mission category. To group institutions at a still higher level, we refer to the two categories in which research is identified as primary to the mission (research essential and balanced) as “research oriented,” and those where teaching is the primary mission (teaching favored and teaching essential) as “teaching oriented.”

These categories avoid the blending of missions that takes place in the Carnegie 2000 classification, allowing us, for example, to combine responses from research-focused master’s and bachelor’s institutions with the bulk of the doctoral institutions. Where the mission categories prove to have greater statistical significance or explanatory power than the Carnegie classifications, we report them accordingly.

### Analysis and Reporting Conventions

We observed the following conventions in analyzing the data and reporting the results:

- Some tables and figures presented in this study report results for fewer than the total number of survey respondents (309). They were adjusted for missing information or to reflect some subset of responses.
- Sums of percentages in some charts and tables may not add up to 100.0% due to rounding.
- The data for each question in the online survey was analyzed for differences in patterns of response among Carnegie classes, Canadian and U.S. institutions, private and public institutions, U.S. region, research/teaching mission, and institutions of varying size. Institution size is determined by the number of full-time-equivalent (FTE) enrollments. We also looked for associations between other combinations of variables as appropriate. Differences that were both meaningful and statistically significant were noted in the text and/or the supporting figures and tables. Note that

<table>
<thead>
<tr>
<th>High-Level Category</th>
<th>Category</th>
<th>Mission</th>
</tr>
</thead>
<tbody>
<tr>
<td>Research oriented</td>
<td>Research essential</td>
<td>Research and teaching are the primary missions, but research is what really drives faculty and institutional success.</td>
</tr>
<tr>
<td></td>
<td>Balanced</td>
<td>Research and teaching are both primary missions, and they are equally important for faculty and institutional success.</td>
</tr>
<tr>
<td>Teaching oriented</td>
<td>Teaching favored</td>
<td>Teaching is the primary mission, but faculty research is rewarded.</td>
</tr>
<tr>
<td></td>
<td>Teaching essential</td>
<td>Teaching is the primary mission, and faculty research does not factor heavily in faculty and institutional success.</td>
</tr>
</tbody>
</table>
a statistically significant relationship between variables does not necessarily indicate a causal relationship.

- The Likert scales used in the online surveys are footnoted in the tables and figures showing results for these survey questions.
- Following what we believe to be the predominant practice among IT professionals, we treat the word “data” as singular rather than plural. Aside from noting that this seems to be the trend in ordinary speech, we believe that IT administrators tend to look at data as a resource or entity to be managed rather than as a collection of datum units. We expect, for example, that not many IT professionals would be inclined to say that “few data are stored on floppy disks anymore, but more and more are stored on key disks,” however much they might agree with the proposition.

**Overview of Respondents**

Invitations to respond to our survey were sent to the ECAR representative or (where the institution was not an ECAR subscriber) to the EDUCAUSE institutional representative at each EDUCAUSE member institution. In most cases, the invitee was the CIO. From 1,733 invited institutions we received 309 responses (a 17.8% response rate). Of the 309 respondents, 290 were from the United States or its territories and 19 were from Canada.

Figure 2-1 shows how respondents’ institutions were distributed among Carnegie categories, alongside the distribution of EDUCAUSE members and the overall population within each category at the time we closed our survey in March 2009. Proportionately, we had the highest participation from doctoral institutions. We also had more respondents from public institutions (58.6%) than from private ones (41.4%). Figure 2-2 shows the distribution

![Figure 2-1. Survey Respondents, by EDUCAUSE Membership and Carnegie Class](image-url)
and Carnegie class composition of the research/teaching mission categories.

The median FTE enrollment of respondent institutions was 4,219, and the mean was 7,642. As Figure 2-3 shows, smaller institutions contributed disproportionately to the results; 70.8% of respondent institutions had FTE enrollments of 8,000 or fewer students.

The survey was completed predominantly (78.4%) by the senior-most IT leader (CIO or equivalent; see Figure 2-4). Another 7.9% were directors of academic or administrative computing; combined with other IT management titles, a total of 96.8% reported IT-related titles. About 7 out of 10 respondents (68.5%) agreed or strongly agreed that they were personally very involved in data management policy and planning at their institution, yielding a median 4.00 and mean 3.68 level of agreement on our 5-point scale (where 1 = strongly disagree and 5 = strongly agree).

**Study Organization**

The rest of this study presents the results of our quantitative survey and the qualitative research that accompanied it.

To set the stage for later analysis, Chapter 3 looks at the institutional context for data management. We assess what respondents told us about trends in the amount of storage they manage and how well they believe their data infrastructure can handle the different varieties of data and content that their institutions need to accommodate. We look as well at drivers and barriers to institutional investment in data management and at the organizational context for data management expertise and improvement.

Chapters 4 through 7 look at different aspects of data management and the infrastructure, policies, and practices brought to bear on them. Chapter 4 examines in detail respondent assessments of enterprise data quality and initiatives to improve it, then looks at institutional analytics capabilities. In
Figure 2-3. Survey Respondent Institutions, by FTE Enrollment (N = 301)

- 1,001–2,000, 23.3%
- 2,001–4,000, 24.9%
- 4,001–8,000, 22.6%
- 8,001–15,000, 13.3%
- 15,001–25,000, 11.0%
- More than 25,000, 5.0%

Figure 2-4. Survey Respondent Titles (N = 305)

- Senior-most IT leader (e.g., CIO or equivalent), 78.4%
- Director of academic computing, 2.3%
- Director of administrative computing, 5.6%
- Other IT management, 10.5%
- Other academic management, 0.7%
- Other administrative management, 0.3%
- Vice president/provost/vice provost or equivalent (non-CIO), 2.3%
- Other academic management, 0.7%
- Other administrative management, 0.3%
Chapter 5, we consider the extent and impact of data stewardship policies and practices and look at the policy environment related to data security, use, and access. Chapter 6 reports on institutional content management environments, looking at the extent of content management solutions and their impact on managing data variety, as well as examining practices related to the institutional website and records management. Chapter 7 looks at the increasingly complex problem of research data management, with a focus on how research data management support is distributed throughout the institution and on how institutions approach long-term research-data preservation issues.

In Chapter 8, we look synthetically at five outcomes measures chosen to represent different aspects of institutional data management, considering how desirable outcomes are related to respondent practices and policies. Finally, Chapter 9 looks at the future of data management in higher education, with an eye on the implications of our study results.

Endnotes
2. 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.