Key Elements of a Data Strategy

Braden J. Hosch, Ph.D.[1]

Introduction

As discussed earlier in this volume, a data strategy is an intentional action plan to capture, integrate, and use data to advance an institution's mission and goals. This chapter seeks to extend beyond theory into practice by describing seven key elements of an effective data strategy: (1) data acquisition, (2) data governance, (3) data quality, (4) data access, (5), data literacy and usage, (6) data extraction and reporting, and (7) data analytics. Many of these can be reconfigured to fit organizational context and maturity, but all must remain present in some form. By incorporating these components in its data strategy, an institution will ensure the availability of sufficient quality data to advance the institution's mission and goals.

The maturity and sophistication of each of these elements are context-specific as well as specific to each major data store or system housing data of value, more broadly termed "data assets." While a fully realized institutional data strategy ultimately encompasses all data created or used by the institution, in practice, the development of a data strategy and the articulation of its key elements require scope-based prioritization. For instance, a college will likely place high priority on its enterprise system for students and faculty, such as Banner or PeopleSoft, and integration with course evaluation platforms and learning management systems, but might delay full integration with systems for advancement and faculty research data to later phases.

Such decisions should be informed by institutional priorities and goals as well as a return-on-investment analysis. One consideration, for example, would be the extent to which

[1] Braden J. Hosch, Ph.D. is the Associate Vice President for Institutional Research, Planning & Effectiveness at Stony Brook University, Stony Brook, NY.

integrating student information systems (course grades, addresses, etc.) with faculty research data stores (data sets for NSF research, archival documents for historical monographs) would produce value for the institution and advance its goals. Since little value would likely be added, faculty research data stores might be placed outside the scope of initial phases of the data strategy.

Table 1. Key Elements of a Data Strategy

Data Acquisition	Data Governance
How the institution obtains its data Build an inventory of data assets. For each one, establish a written plan for: Identification Prioritization Capture Storage Linkage Curation	How people make decisions and behave with respect to how data will be defined, produced, used, stored, and destroyed Establish: Decision-making body and rules Data dictionaries Data stewards
Data Quality	**Data Access**
How data will be maintained to be complete, valid, consistent, timely, and accurate to make them appropriate for a specific use	How authorized individuals can obtain and use data while maintaining privacy and security Establish written plans for: Accessibility Security

Data Usage & Literacy	Data Extraction & Reporting	Data Analytics
How data users understand and use data Establish: Data user responsibilities Training/education protocols Usage metrics	How data will be queried and retrieved from storage and delivered to users Establish protocols for: Extraction Reporting	How data will be used through dynamic and visual deployment for benchmarking, exploratory and causal analysis, and prediction and forecasting

Data Vision

Much like a vision statement is a forward-looking articulation of what an organization would like to become, a data vision statement indicates how data will help an organization

realize its mission and strategic goals. Importantly, the data vision should not focus on IT but rather highlight how data can benefit people and operations. The British Library (2017), for instance, has formulated this strong and straightforward statement: “Our vision for the British Library is that data are as integrated into our collections, research and services as text is today.”

Colleges and universities should consider how such a statement supports their context, mission, goals, and resources. An aggressive approach might be to advocate integrating the university’s data assets in the same way that devices are networked today. Articulating a vision of this scope would highlight the need to deploy significant resources for its realization, much like IT has invested in networking staff, infrastructure, policies, and risk management. Alternatively, the vision might be less resource intensive but still forward looking. For instance, the statement might emphasize how people interact with data: Students, faculty, and staff will use data assets the way that they use email today. The difference in focus between these two examples illustrates the importance of deliberating about an organization’s data vision; the future it articulates will drive decision making, resource allocation, and prioritization. A clear statement of data vision will guide how an institution approaches the subsequent components of its data strategy.

Key Elements of a Data Strategy

Even though the concept of an organizational data strategy is a relatively new development in industry and only nascent in higher education, some models have been advanced to articulate the basic components and approaches. Levy (2018) breaks down an organizational data strategy into five core actions in an organizational data strategy: (1) identification of data and its meaning, (2) storage of data in persistent structures, (3) provisioning of data to make it re-usable, (4) processing of data to combine data residing in disparate systems, and (5) governance

of data to promote effective usage. In another approach, Caruthers and Jackson (2018) suggest that new chief data officers develop an immediate data strategy encompassing six components: (1) stability and rationalization, (2) data culture and governance, (3) existing IT initiatives, (4) data exploitation and integration, (5) data performance and quality, and (6) data security. The target data strategy described subsequently is more of a guide for organizational change management than an action plan.

While these approaches offer some useful guidance, they generally do not provide sufficient detail about what needs to be considered and done to craft and implement a data strategy in the context of higher education. This chapter elaborates on a data strategy framework developed by Stony Brook University, an internationally ranked research institution (Hosch, 2017). The seven components are described in greater detail herein, offering more direction for how to develop and launch a data strategy at any college or university. Notably, these components may differ among data assets depending on their priority and relationship to institutional goals and operational needs.

1. Data Acquisition

Simply put, data acquisition is how an institution of higher education obtains its data. Employees and students generate data internally, and data also comes from outside sources. Colleges and universities already acquire data through the admissions process, teaching and learning, and administrative operations, of course, but the process for doing so has in general been reactive and unsystematic. For instance, the admissions office at the request of leadership may conduct a competitive analysis and so obtain data from the National Student Clearinghouse (NSC) about where non-enrolling applicants eventually matriculated, but NSC data may not be stored or regularly re-used. The financial aid office receives reports from the federal government

about the re-payment status of former students who borrowed educational loans, but the data may be stored in Excel files on a desktop computer. The advancement office may maintain all of its records in a separate system for donor management. In all of these instances, the institution has acquired data, but absent a formal data strategy, these data assets will not be leveraged to their fullest extent. An effective data acquisition strategy involves six activities.

Activity 1.1: Identification

An early step in formulating a data strategy is to establish and maintain an inventory of data assets and assess the maturity of acquisition processes. Kiron (2017) documents the importance of a deliberately constructed and managed data inventory. This inventory may be large and will grow. Data in the enterprise system, such as PeopleSoft or Banner, is the most obvious data asset, but in addition to this important system and the examples above, other assets include the learning management system (LMS); comparative data, including those from IPEDS, ranking publications, and other sources; data in vendor-based systems, such as those for recruitment, student success, assessment, or space management; faculty activity data; library data systems; the data warehouse(s); residence hall management systems; facilities access and energy usage data; network usage data; survey systems; and document imaging repositories. The inventory may also cover external data assets that offer insight into the internal and external environment; these include data feeds from social media, labor market demand and outcomes, and other real-time data that may inform strategic decision-making. At a minimum, the inventory should include the name of the asset, a brief description of the data it houses, the vendor (if applicable), the unit and person responsible for it, and the storage location of the data (university server, office file share, vendor cloud, etc.).

Activity 1.2: Prioritization

Once an inventory is well populated, the next step is to establish a process to prioritize integration into the data infrastructure. Just as an effective IT governance system includes an agreed-upon process for prioritizing technology projects to allocate resources to meet the most important organizational goals and needs (Weill & Ross, 2004), an effective data strategy delineates which data assets in the inventory should receive attention first. The institution's mission, goals, and strategic objectives should guide the prioritization process. For instance, a university strategically targeting improvements in undergraduate student success may place priority on the ERP system and the LMS system to understand how progress through course-level experiences informs degree progress, while deferring action on data assets for advancement and research. The personnel involved in setting priorities for data integration should understand the strategic goals of the institution as well as the potential for various data assets to advance these priorities.

Activity 1.3: Capture

For each data asset, the data strategy should identify current and optimal capture procedures. Data may be manually keyed by employees or students; optically scanned from documents or barcodes; and/or manually or automatically imported from other sources. In practice, many systems use multiple data capture methods, For instance, undergraduate applications data may be sent to an institution from the Common Application through a direct feed, while medical records are optically scanned by staff in the health center, and current address and contact information is manually keyed by each student. As systems of capture mature, they rely more on an application programming interface (API) that feeds data directly to an institutional system and do not require the attention of employees except to ensure that they

execute as scheduled. More mature systems also re-use data by transporting information that has already been captured in one system into another that requires the same entries.

Activity 1.4: Storage

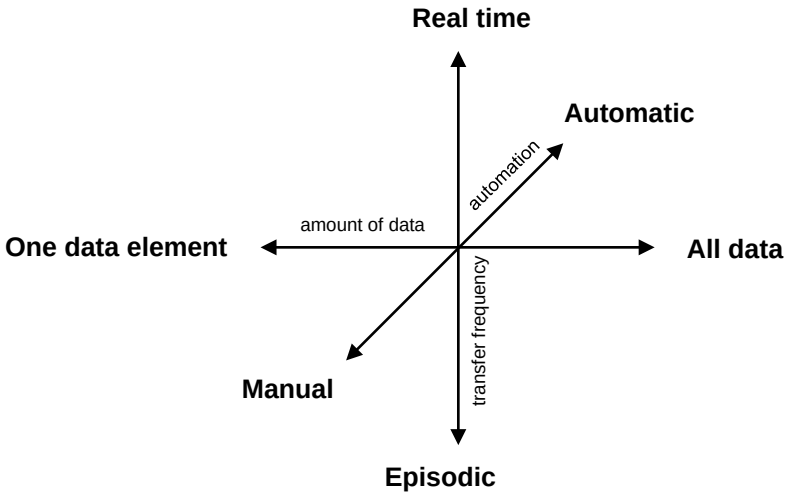
The data strategy should identify current and optimal storage areas for each data asset. The current locations of the enterprise system for student records, the LMS, and other major systems are well known to IT professionals and can be identified easily. Data assets maintained by individual units may be less well-known or defined. For example, the institutional research (IR) office may manage its own warehouse or storage system for IPEDS unit records and benchmark data, and it may preserve external data sourced from ranking publications or environmental scans only on an ad hoc basis in a file share. In instances like these, the current location of the IR file share should be recorded and an optimal location identified; possibilities include the institutional data warehouse, a data lake, or a location where a federated analytics system could access the file share and integrate it with other data assets. As a result of the prioritization process, the optimal location will not always be the most integrated option for storage, but rather the option that will best advance organizational goals in an environment where resources are limited. Considerations about access may influence where data may be stored, as in the case of financial aid records received directly from the Free Application for Federal Student Aid (FAFSA), restricted-use license data for federal sample surveys, and other data assets that are subject to data use agreements.

Activity 1.5: Linkage

The data strategy should identify current and optimal procedures to connect each data asset to others belonging to the institution. Articulation of how data assets do and should connect with each other is context dependent, based upon existing organizational architecture and

priorities. A range of current and optimal approaches to linkage will emerge, from a manual Open Database Connectivity (ODBC) connection between a file share and a desktop workspace to automated extract, load, and transform (ETL) procedures that move data in real time into a data lake or structured repository. Specifics about how data assets are connected should include the amount of data involved, the frequency of transfer, and the level of automation. As illustrated in Figure 2, dimensions of automation, frequency of transfer, and data quantity exist on a multi-dimensional spectrum. As connections require more data elements, more automation, and more currency, more monetary and IT resources will be required to implement and maintain them.

Figure 1. Dimensions for Data Linkages



For example, a residence life operation that runs on a separate housing management system may push account charges and local addresses to the enterprise system on a nightly basis and pull demographic data back into the housing system using an API. Such a linkage constitutes a daily two-way automated transfer of limited data elements, and may be perfectly sufficient to conduct business. However, if institutional goals to improve student success mean prioritizing an examination of the effects of roommates and room changes on academic performance, then additional data elements may need to be pushed to the enterprise system or other systems to

determine the relative importance of these factors. The optimal state may thus involve increasing the amount of data flowing into the enterprise system or constructing an analytical architecture that can connect to the entirety of both systems in real time. In advancing a data strategy, it is important to resist the impulse to immediately connect everything to everything else and instead prioritize those connections that will produce early returns. A mature data architecture will integrate a majority of administrative data assets and will illustrate how reporting and analytics will connect to these assets (Campbell, Smith & Kumar, 2018).

Activity 1.6: Curation

For each data asset, the data strategy should identify how data will be updated and maintained to preserve value. Data elements do not maintain themselves; a group of professionals and processes maintain them to ensure that users receive high-quality data. The data strategy should identify who is responsible for updating and maintaining data. It is further useful to articulate what systems are in place to ensure the appropriate delivery of data. This may involve a series of error checks or audits or an automated notification of a completed or uncompleted process. It may also involve more detailed business analysis and communication with vendors or data providers to understand changes in data acquisition, including data feeds, definitions, cycle time, measurement point or other salient information.

2. Data Governance

Data governance formalizes behavior around the definition, production, storage, usage, and destruction of data to enable and enhance organizational effectiveness. Importantly, data governance is about people and business processes more than it is about data, and while IT professionals should participate in data governance, it should not be relegated to or led by an institution's IT unit. Further, development of effective data governance is a sufficiently complex

initiative that has received extensive treatment by multiple authors (Bhansali, 2014; Seiner, 2014), and our discussion here is necessarily limited. That said, some essential elements of data governance deserve attention. Otto (2011) notably identifies three characteristics of data governance systems: 1) connection to the organization's formal and functional goals, 2) decision-making rights, and 3) roles and committees. It is also significant that the "formal" nature of data governance requires that all three of these features be documented in written form and preferably made broadly available throughout the organization, such as by posting them on a website or intranet.

In translating these characteristics to a higher education data strategy, the data governance approach should be articulated for each data asset to include 1) a designated decision-making body with established rules for how it makes decisions about data; 2) individuals to provide data stewardship for various assets; and 3) a system that produces and maintains formal (written) data dictionaries that store metadata. Colleges and universities may deem it desirable to establish a data governance system that encompasses all their data assets, although the complexity of any given institution may render a unified system impractical. For instance, a university with significant medical and hospital services may wish to govern patient data protected by the Health Insurance Portability and Accountability Act (HIPAA) using specialized structures and policies, rather than placing those data elements under the general purview of an institutional data governance council.

Data stewards play an essential role in data governance, although many colleges and universities have no formal descriptions or written sets of expectations, activities, or deliverables for these positions. Data stewards conduct the day-to-day business of data governance and are accountable for effective control and use of data (Plotkin, 2014; Knight, 2017). Plotkin identifies

five types of data steward roles: domain data stewards, business data stewards, technical data stewards, operational data stewards, and project data stewards. Business data stewards are accountable for data within a particular area, such as a college or school run by a dean; they work with stakeholders to make recommendations on data issues, manage metadata for their area, champion stewardship, and communicate important information back to data users in their areas. Domain data stewards are responsible for widely shared areas of an institution, such as the registrar for university records or the controller for financial data; these individuals work with business data stewards to build consensus and consistency across the domain. Technical data stewards are usually IT staff; they provide expertise on applications, ETL, data stores, and other links in the information chain and are assigned by IT leadership to support data governance. Operational data stewards provide support to business data stewards and hold campus roles like department course scheduler or unit hiring manager. They help enforce business rules for the data they use and may remediate data under their purview when needed; they may recommend changes to improve data quality. Project data stewards are less common in higher education but may be appointed to help domain data stewards implement a specific project, such as a course scheduling system or degree audit system.

Data stewards protect and curate the value of data assets under their purview. Specifically, they oversee management of selected data assets; participate in data governance and carry out decisions; assist in creating and maintaining data dictionaries and metadata; document and update rules, standards, and procedures relevant to their area of responsibility; ensure data quality and manage specific issues; communicate appropriate use and changes; and manage access and security (Stanford, 2012). These responsibilities are not trivial, and stating them

explicitly in job descriptions and performance expectations can help ensure that they are properly valued and carried out.

The responsibility for maintaining data dictionaries carries particular importance because effective use of data requires a shared "common understanding of the meaning and descriptive characteristics of that data" (International Standards Organization, 2004). An organization may be streamlined enough to adopt a systematic master data management (MDM) protocol that centralizes definitions of all data elements in a master data dictionary. But the proliferation of data assets, including those from third parties, may require the data governance body to set data dictionary standards and distribute management of the dictionaries themselves to data asset managers. For instance, Stony Brook University's data dictionary standards include a set of principles and required elements (data store, table name, data element, data element name, definition, source and data logic, data type and length, allowable values/parameters, semantic rules, data steward, date created, and date updated), with specific directions and examples for how to manage the data:

Data Dictionary Principles

1. Data dictionaries are designed to promote communication and production of meaning; as such dictionaries document the existence, meaning, and use of data elements
2. Data dictionaries must be accessible to all users who enter and extract data from a data store
3. Data stewards must actively maintain data dictionary contents, including definitions, values, and other metadata
4. Data caretakers and users are responsible for actively using data dictionaries to correctly enter, select, and analyze data elements

5. Data dictionaries should be reviewed on a regular schedule to ensure currency (Stony Brook University, 2017)

3. Data Quality

In many organizations, calls for data governance and formulation of a data strategy are typically prompted by complaints about data quality. The data quality problem is often manifested when analysts produce different answers to the same question, when invalid values appear in reports, or when analysts spend inordinate amounts of time manually cleaning data before issuing reports. In some cases, these issues are definitional and can be addressed through data governance, but in others, the data sources are the culprit because of missing elements or logically impossible values populating fields, such as a Connecticut address in a New York county or an undergraduate art major housed in the law school. Statistics Canada (2002) offers a useful definition of data quality as "the state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use."

Missing data elements complicate even the simplest analytical work and can make it difficult for audiences to interpret percentages and ultimately understand findings. Further, requestor-based projects can be derailed by missing data; missing ZIP codes may prevent proper attribution of students to legislative districts and weaken advocacy efforts, or drop-out risk may be improperly modeled because of missing student activity information. Invalid or illegal values in data fields likewise complicate analysis because they require ad hoc data cleaning by the end user that is not replicable and entails a high probability of error. For all intents and purposes, Data elements that are unavailable when they are needed are just an extreme case of missing data, and inaccuracies in data pose obvious problems.

Through an analysis of various data quality approaches, Batini et al. (2009) found that successful methods incorporate a collection of contextual information about business processes and storage practices, an assessment of the quality of extant data, and an improvement process. The contextual phase of data quality management includes understanding how business units operate, store data, and face challenges and costs arising from data quality issues. The data quality assessment phase requires a comparison of existing data to reference values from data dictionaries as well as a discussion with stakeholders about the most critical areas for attention.

In organizations where data quality management is nascent or developing, these assessments are ad hoc and reactive. In more mature data quality management systems, quality assurance measurement happens automatically and initiates reports highlighting data to be corrected. In the most advanced systems, organizations have distinct measurements of data quality across all data assets, with measures of completeness and validity across all elements, and also have implemented strong validation procedures upon acquisition of data to minimize capture of invalid data. Approaches that incorporate error reporting can generally be developed locally, but systemic measurement of data quality across all data elements and sources generally requires a dedicated software solution or application.

The data quality improvement phase involves multiple activities, which are, in part, a function of the maturity of the institution's data management practices. Batini et al. (2009) found that organizations manage data quality effectively when they evaluate the costs of remediation, assign process and data responsibilities, identify causes of errors and appropriate remediation strategies, design and implement process controls, and monitor improvements. In terms of costs, institutions may have received lists of co-curricular high school activities from the Common Application as free-response text fields, with reported items such as "dance," "dance team,"

“competitive dance,” “clogging,” and “Irish step dancing,” not to mention variations with misspellings and errant punctuation. Data cleanup of this field could be accomplished in various ways, but the costs might outweigh the benefits.

The data governance system should assign data stewards to all data assets to ensure quality control. Identification of causes of errors and appropriate remediation can be difficult to automate; such activities typically require an analyst who understands business processes and can troubleshoot why data stores reflect unexpected results. Monitoring of improvements can be done on an ad hoc basis, but in organizations with more mature data quality management processes, direct measurement of all data against reference values from dictionaries can offer compelling metrics about the effectiveness of any improvement efforts. Further, these measurement practices generally rely upon software solutions to determine the extent to which all data fields are complete and meet parameters in the dictionary, as well as to cross-validate them with other data elements.

4. Data Access

An institution’s data strategy should establish provisions for data access 1) to ensure accessibility: allowing authorized individuals to obtain and use data when and where necessary and 2) to provide security: protecting privacy and preventing unauthorized use of sensitive information. Moreover, a comprehensive data strategy will tailor accessibility and security requirements to each data asset, although establishment of an overarching framework to classify data assets for access and security protocols is helpful for streamlining purposes.

4.1. Accessibility

Data access in a closed paradigm is often conceptualized as user authentication to ensure that only authorized individuals have access to sensitive or restricted data; this principle is

covered below under “Security.” Conversely, accessibility in an open paradigm extends data out to these authorized users so that they have data when and where they need it. Considerations include the devices and networks on which data may be accessed, the applications that may be used to work with the data, and the timeliness of data. Institutional needs must be balanced with security demands as well as potential return on investment. For instance, if a college president wishes to access her executive dashboard via her tablet using a 4G network, then use of unit record data with personally identifiable information (e.g., student IDs or grades) as the architectural foundation of the dashboard may pose security concerns. In another instance, an enrollment manager may want real-time access to registration numbers to monitor progress toward enrollment and retention goals but may have to settle for receiving the figures nightly via the data warehouse if the cost or security limitations of obtaining them directly from the live student information system are too significant. The data strategy should balance the data vision and institutional goals against costs and potential return. Further, since security restrictions may place additional limits on some data assets but not others, accessibility may be asset specific.

4.2. Security

Data security in the data strategy should incorporate both an institutional approach and an asset-based approach. At the institutional level, colleges and universities already have to comply with federal laws pertaining to the handling of education records under the Family Education Rights and Privacy Act (FERPA), financial records under the Gramm-Leach-Bliley Act (GLBA), and potentially health records under the Health Insurance Portability and Accountability Act (HIPAA). GLBA in particular has provisions requiring institutions to maintain a formal information security program and designate an employee to coordinate it. In 2016, the Department of Education issued a Dear Colleague Letter (GEN-16-12) reminding institutions

that compliance with GLBA is a requirement of their Program Participation Agreements in Title IV student aid programs and also strongly encouraging institutions to adopt protocols outlined in NIST SP 800-171 for protecting "controlled unclassified information" (Ross et al., 2015). At the institutional level, the data security program should certainly comply with GLBA, although for many colleges and universities the protocols set forth in NIST SP 800-171 will be aspirational. From a data strategy perspective, institutions need to establish security policies that classify sensitive information, specify responsibilities of users, and include provisions for authorization protocols. The following are two strong examples:

Stanford University

The Administrative Guide's chapter on computing contains specific policies on information security, including information sensitivity, stewardship, access, and authentication protocols. See especially subchapters 6.3.1 (Information Security) and 6.4.1 (Identification and Authentication Systems).

University of Michigan

Safely Use Sensitive Data is a website that details data classification levels, methods to protect sensitive data, and what kinds of data are allowable in various university systems.

Many institutions have established a chief information security officer (CISO) position reporting to the chief information officer or at times to the president or the board, but the organization and staffing models are less important than ensuring that the function is carried out and that sufficient policies are established (Pomerantz & Grana, 2017). Additionally, the General Data Protection Regulation issued by the European Union has prompted many colleges and universities to establish more intentional and unified privacy policies. Data security and

individual privacy are related but distinct concepts, and larger institutions may wish to consider the utility of a chief privacy officer as personal data proliferate.

The mere existence of institutional policies that establish levels of information sensitivity and access protocols is not enough, however; each data asset must be assigned these classifications and protocols. The University of Michigan's practice of centrally maintaining them and communicating them via a website is a leading example of how to accomplish this. A sound data strategy will also incorporate these protocols beyond centrally managed assets to those managed by units or federated into analytical networks. Restrictions on some federal financial aid data, such as information reported to institutions via FAFSA, for instance, prevent linking many of these data elements with more broadly accessible analytical databases. Therefore, institutions must take care to ensure that appropriate security and privacy protocols are maintained for each data asset and element while pursuing a strategy of broader data integration.

5. Data Literacy and Usage

An institutional data strategy should establish a plan to ensure that the people who regularly work with data understand what it means, can explain both its proper uses and its limitations, and can use it to support decision-making and to improve operational effectiveness. This aspect of the data strategy involves setting competencies and providing sufficient professional development and on-board explanations to ensure that users are able to use data appropriately. Organizations are increasingly finding that employees do not have the data skills they need (Harris, 2012; Bradford, 2018). Especially in larger distributed environments with broad access to "democratized data" through analytics, these systems for data literacy have to be

web-based to scale them sufficiently. Again, they may also need to be specific to the data asset, as in the case of the University of Washington Business Intelligence Portal Tour (2018).

Additionally, formalized policies can assist with advancing data literacy, for example by requiring job postings and position descriptions to include relevant data competencies. Establishing a formal policy for user responsibilities can also be valuable; a statement in the Stony Brook University Data Governance Framework (2016) establishes user responsibilities to "recognize that institutional data and information derived from it are potentially complex," include source information when distributing data, guard against potential misinterpretation, respect individual privacy, maintain security standards, and report data quality issues to data stewards.

Closing the loop to measure usage should start with metrics of report usage and access, but should not end there. The most basic way to gauge usage is often simply to review which reports are accessed the most. Valuable information can be gleaned about what is working in an analytics system and, perhaps more importantly, what is not. When a report or dashboard goes unused, it is often because users either don't know the tool exists or don't know how to interpret the data contained therein. In some instances, redesign and better communication can solve these issues, but it may well be that users need more training, especially with closing the loop. Ransbotham, Kiron, and Prentice (2015) found an increasing gap between the sophistication of analytics and the ability of managers to interpret and use the results. Indeed, the largest corporate challenge is not producing results but rather translating results into action. More sophisticated measures of how users understand and put data to use can take the form of short follow-up surveys asking two or three straightforward questions, such as "To what extent did the data

delivered meet your needs?" (Likert scale) or "What decision or action was made based on the data?" (short free response).

6. Data Extraction and Reporting

While the acquisition provision of the data strategy covers how to get data into institutional systems, the extraction and reporting component formalizes how to query and retrieve data from storage and deliver it to users through both regular and ad hoc reporting to support day-to-day operations. Methods for querying and extracting data from storage should be identified, along with user types associated with each extraction method. The data strategy should establish roles for users who access raw data, build reports, or simply access reports.

Some personnel, such as financial aid and institutional research staff, will need direct access to data storage areas to extract large data sets with different parameters. This level of direct access through a virtual private network (VPN), secure file transfer protocol (SFTP), or other secure transfer protocol will need to be planned and established. Staff in IT or business intelligence (BI) units who build reports for other campus users will need similar access as well as a means to deliver reports securely to these constituencies. Finally, general campus users will require a way to access these reports. For established university systems, such as the student records system or financial system, these protocols are likely already in place. A university will need to consider how to incorporate similar extraction protocols for all of the other data assets in its inventory. In most cases, the ways in which data may be extracted from the student records system will differ substantively from those used for the learning management system (LMS), the faculty information system, or even the data warehouse. The data strategy should articulate these differences and assess the value and intentionality of each system of data extraction.

Reporting, which is distinct from analytics (below), is “the process of organizing data into informational summaries in order to monitor how different areas of a business are performing” (Dykes, 2010). Reporting represents a basic operational function and often involves lists or very simple statistics, such as counts and averages. Class rosters, lists of students on probation or suspension, daily counts of students registering before the start of the term, and statistics produced for compliance purposes such as IPEDS are all examples of the kinds of reports that systems should be designed to prepare.

The data strategy should establish principles to guide the unit or units that build and deliver reports, including those that may assemble data from distinct data assets. Among these should be a provision that reports must support operational objectives. That is, a report should be designed to accomplish a specific task, and this task should be clearly stated in the report. For example, a class roster could include a description such as “This class roster is designed for instructors to know which students are officially registered for their class section as of the date the roster is prepared. Please report discrepancies to the Registrar.”

Additionally, the data strategy should establish a searchable inventory of reports and their intended use, and the inventory should be maintained in an accessible area. Reports should be automated based on return on investment that includes some forecasting about the stability of the environment or reporting needs. Finally, the data strategy should institute some metrics for report usage, such as how often each report is run and how many distinct users run a report in a given time period. Just as with data literacy, occasional user surveys to assess what decisions are made based on reports can provide invaluable insight into how much reports contribute to the accomplishment of institutional objectives.

7. Data Analytics

Analytics describe the past (descriptive analytics), explain the present (exploratory analytics), forecast the future (predictive analytics), and propose future courses of action (prescriptive analytics). Data analytics are the tools and output of data analysis. In many ways, such tools have always been employed by colleges and universities. However, advances in computing power and the proliferation of digital records have led to rapid development in analytics. Contemporary analytics offer speed, ease of use, interactivity, and utility for decision-making. Moreover, they increasingly include machine learning approaches and deployment of artificial intelligence. Analytics have also been integrated into data systems, so that room scheduling systems, donor management systems, faculty information systems, and the like all include their own data analytics using some level of visualization and forecasting. This reality of multiple analytics systems native to specific data assets renders the development of a coherent analytics strategy complex.

The data strategy should establish a plan for data analytics for each data asset and for the institution as a whole and set priorities for the integration of data assets into the institutional analytics system. The data strategy should acknowledge that a successful analytics system requires maturity in other aspects of the data strategy, including data acquisition, governance, quality, access, usage, and extraction. The analytics integration should be designed with key questions in mind that will advance the institution's strategic priorities; these often involve revenue generation, cost containment, and risk mitigation. Examples of common descriptive and exploratory analyses appear in Figure 3. The data strategy should identify preferred analytics systems for particular questions or analyses. For instance, the analytics native to the faculty information system may be the preferred source for understanding faculty research productivity

and impact but should not be used as a source for official counts of faculty members and their demographics.

Table 2. Examples of Analytics Outputs

Descriptive/ Exploratory	**Revenue Generation**	**Cost Containment**	**Risk Mitigation**
	Revenue per credit hour Revenue per program Application pool analysis Revenue per recruitment event	Cost per credit hour Class size distribution and optimization Space utilization Faculty workload Unit staffing per service levels Cost per service delivery	Compliance monitoring Budgeted vs. actual expenses Cost as percent of family income
Predictive/ Prescriptive	**Student Success Factors**		**Administrative**
	Individual predictions for: - Student grade point average - Likelihood of on-time graduation - Likelihood of attrition - Likelihood to register for a class Individual prescriptions for: - Low-impact interventions (nudges) - Medium-/high-impact interventions		Faculty progress toward tenure Forecast of research productivity Forecast of department/program ranking Course demand Enrollment projections Revenue/cost forecasts

Data Tools

While recognizing that multiple data tools will likely play a part in its overall execution, the data strategy should identify those preferred for specific institutional functions. Acceptance of multiple tools is necessary since 1) new tools are developed increasingly rapidly, driving innovation and reducing acquisition costs; 2) users prefer specific products and have tool-specific expertise; 3) some data assets require use of specific tools; and 4) the relative strengths and weaknesses of various data tools should prompt users to select the tool that best matches the requirements at hand. For instance, a small college may find that it can limit statistical applications to a single package such as SAS, but mid-size and larger institutions will almost certainly need additional applications, including SPSS, R, and Stata, to support academic and research missions as well as to leverage the application-specific expertise of different analysts. Tool selection should generally follow identification of key functional questions and an

assessment of local resources for deployment and analysis. Considerations should include speed, ease of use, capabilities for analysis, support for training and professional development among employees, deployment on premises or in the cloud, and capacities to access to multiple data assets.

Data tools or services are available to support any of the seven key elements of the data strategy described in this chapter. Four areas deserve special attention here: 1) storage and integration, 2) reporting, 3) business intelligence, and 4) advanced analytics. These four functions exist on a continuum moving data from various sources and transform the data into intelligence, and many tools include components of each, with some vendors offering full vertical integration of all four functions. Further, as the marketplace for these applications and services is evolving rapidly, the examples provided here are necessarily a snapshot of the market at a point in time. Storage and integration tools may employ traditional data warehousing with extract-transform-load (ETL) procedures into a server environment on premises, such as Microsoft SQL Server. Alternatively, many larger institutions have begun to develop a data lake following an extract-load-transform (ELT) model from source data to the cloud, using Amazon Web Services, IBM Cloud, or Google Cloud Platform (Campbell, Smith & Kumar, 2018; Aldrich, 2018). Such implementations are typically resource intensive, and the data strategy should outline the institutional approach to this level of architecture.

By and large, reporting tools act as a layer on top of the enterprise warehouse, data marts, or data lake, and again typically entail an institutional decision to support a single or limited number of approaches for data extraction and reporting. MS SQL Server Reporting Services, Crystal Reports, and Cognos are all commonly used reporting applications. BI applications have become widespread, and the market is in flux. Leaders in higher education include Tableau,

Microsoft BI, and Oracle OBIEE. As Baier et al. (2018) observe in an evaluation of 20 BI platforms, Tableau provides a user-friendly platform that lowers barriers to adoption among analysts, while Microsoft's platform requires more technical skill but is more economical.

Advanced analytics involve computing approaches that generally exceed the capabilities of BI tools. Python, R, and other base tools may play a role in the analytics space, and may be integrated into more specialized applications, such as Rapid Insight, SAS Enterprise Miner, SPSS Data Modeler, or KNIME. These applications typically conduct the computationally intensive work of prediction, simulation, and forecasting, with results being disseminated either through these tools or through the BI tool layer.

Final Thoughts

Developing a data strategy is a daunting task, but articulating plans for each key element listed here can be a fruitful approach. It must be remembered that creating and implementing a data strategy is not an IT project but rather a systemic process for an organization. Development of the strategy should be inclusive, not only to harness institutional priorities and knowledge but also to garner buy-in from key constituencies. As it evolves, the strategy should be articulated formally and published, at least for institutional consumption. If it is not written down and communicated, then it is not a strategy, it is a secret. Finally, institutions must anticipate the need to devote resources to the data strategy and infrastructure. Sustainable systemic processes that transform an organization generally cannot be supported as distributed assignments on top of existing personnel and functions.

The importance of adopting a data strategy likely cannot be overestimated. Gartner (2017) anticipates that by the early 2020s equity analysts will consider the scope and quality of an organization's data and information holdings to assign a valuation to the company. Higher

education institutions already compete in this environment for students, dollars, time, and political and social positioning. The institutions that most effectively and intentionally adapt to the new reality of pervasive data will establish a competitive advantage through strategic curation and leveraging of data capital that will create opportunities not unlike financial capital does today.

Discussion questions

1. List the major data assets of your organization. To what extent are they integrated with each other, and how does this reflect current institutional priorities and future institutional needs?
2. Recognizing that the formulation of a data strategy should not fall solely to IT but should instead be led by functional operations, who should be involved in developing the data strategy and monitoring its effectiveness?
3. How will the effectiveness of the data strategy be assessed? What metrics should be established for each part of the strategy, and who will be responsible for collecting and evaluating them?
4. How do data security considerations change as more people are involved in creating a and implementing a data strategy?
5. Consider the maturity of your institution in each of the key elements in a data strategy. Which is the element that your institution needs to focus on the most and why? What challenges are you anticipating with the element and what are strategies to mitigate those challenges?

References

Aldridge, B. (2018). “Lead the charge: CSU’s transformational data program,” Educause Annual Conference, Denver, CO. Retrieved Nov. 28, 2018 from https://events.educause.edu/annual-conference/2018/agenda/lead-the-charge-transformational-data-leadership.

Baier L., et al. (2018). "BARC score enterprise BI and analytics platforms.” Retrieved Nov. 28, 2018 from https://media.bitpipe.com/io_14x/io_144396/item_1786174/2018-07-16-barc_score_enterprise_bi_and_analytics_platforms_fin_ASL12447USEN.pdf.

Batini, C., et al. (2009) "Methodologies for data quality assessment and improvement," *ACM Computing Surveys 41*(3) Article 16.

Bhansali, N., ed. (2014). *Data governance: creating value from information assets.* Boca Raton : CRC Press, Taylor & Francis Group.

British Library (2017). *Data strategy 2017*. Retrieved July 16, 2018 from https://blogs.bl.uk/files/britishlibrarydatastrategyoutline.pdf.

Bradford, L. (2018, Oct. 11). “Why all employees need data skills in 2019 (and beyond),” *Forbes.* Retrieved March 18, 2019 from https://www.forbes.com/sites/laurencebradford/2018/10/11/why-all-employees-need-data-skills-in-2019-and-beyond/#1ea80b65510f.

Carruthers, C. and Jackson, P. (2018). *The chief data officer’s playbook*. London: Facet.

Campbell, J., Smith, K. and Kumar, T. (2018). “Building and analytics infrastructure in-house,” Association of Public and Land-Grant Universities Annual Meeting, New Orleans, LA. Retrieved November 24, 2018 from http://www.aplu.org/members/commissions/information-measurement-analysis/cima-presentations-2018/In-House%20data%20infastructure.pdf

Dykes, B. (2010). “Reporting vs. analysis: What’s the difference?” Retrieved October 5, 2018 from https://theblog.adobe.com/reporting-vs-analysis-whats-the-difference/.

Harris, J. (2012) “Data is useless without the skills to analyze it,” *Harvard Business Review*. Retrieved Oct. 1, 2018 from https://hbr.org/2012/09/data-is-useless-without-the-skills.

Hosch, B. (2017). Beyond data governance to data strategy. Annual Forum of the Association for Institutional Research, Washington, DC.

Kiron, D. (2017). “Lessons from becoming a data-driven organization,” *MIT Sloan Management Review 58*(2), 3-13.

Knight, M. (2017). “What is data stewardship?” Retrieved October 15, 2018 from http://www.dataversity.net/what-is-data-stewardship.

Levy, E. (2018). “5 essential components of a data strategy,” SAS Whitepaper. Retrieved Oct. 1, 2018 from https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/5-essential-components-of-data-strategy-108109.pdf.

Otto, B. (2011). “A morphology of the organisation of data governance,” *ECIS 2011 Proceedings* 272. http://aisel.aisnet.org/ecis2011/272

Plotkin, D. (2014). *Data stewardship: An actionable guide to effective data management.* Morgan Kaufmann, Waltham, MA.

Pomerantz, J. and Grama, J. (2017). *IT leadership in higher education, 2016: The chief information security officer*. Research report. Louisville, CO: ECAR. Retrieved Sept. 13, 2018 from https://library.educause.edu/~/media/files/library/2017/7/ers1702ciso.pdf.

Ross, R. et al. (2015). “Protecting controlled unclassified information in nonfederal information systems and organizations,” National Institutes of Standards and Technology Special Publication 800-171. Retrieved Sept. 13, 2018 from: http://dx.doi.org/10.6028/NIST.SP.800-171.

Seiner, R. S. (2014) *Non-invasive data governance: The path of least resistance and greatest success*. Basking Ridge, NJ: Technics Publications.

Stanford University (2012). "Responsibilities related to data stewardship," Retrieved Oct. 15 from http://www.stanford.edu/dept/pres-provost/cgi-bin/dg/wordpress/wp-content/uploads/2012/01/DG-Data-Stewardship.doc.

Stanford University (2017). *Administrative guide: Chapter 6 Computing*. Retrieved Sept. 13, 2018 from https://adminguide.stanford.edu/chapter-6.

Statistics Canada (2002). *Statistics Canada's quality assurance framework - 2002*. Statistics Canada Catalogue no. 12-586-XIE.

Stony Brook University (2016). *Data governance framework at Stony Brook University*. Retrieved October 1, 2018 from https://www.stonybrook.edu/commcms/irpe/about/_files/DataGovFramework.pdf.

Stony Brook University (2017). *Data dictionary standards*. Retrieved Sept. 6, 2018 from https://www.stonybrook.edu/commcms/irpe/about/data_governance/_files/DataDictionaryStandards.pdf.

Ransbotham, S., Kiron, D., and Prentice, P. (2015). "Minding the analytics gap," *MIT Sloan Management Review 56*, 63-68.

University of Michigan (2018). "Safely use sensitive data," retrieved Sept. 13, 2018 from https://www.safecomputing.umich.edu/protect-the-u/safely-use-sensitive-data.

University of Washington (2018). "BI portal tour," retrieved Oct. 1, 2018 from http://itconnect.uw.edu/work/data/bi-portal-intro.

Weill, P. and Ross, J. (2004). *IT governance: How top performers manage IT decision rights for superior results*. Harvard Business School Press, Boston, MA.

Widmer, L. (2014) "Dealing with the data." *National Underwriter Life & Health Breaking News*, 25 Feb. 2014. *Business Economic and Theory Collection*.