POLICY FORUM

HIGHER EDUCATION

Data blind: Universities lag in capturing and exploiting data

Study finds a pervasive void of infrastructure thinking

By **Christine L. Borgman**[1] *and* **Amy Brand**[2]

Research universities are large, complex organizations that generate vast amounts of administrative and research data. If exploited effectively, these data can aid in addressing myriad challenges. Yet universities lag behind industry, business, and government in deriving strategic value from their data resources (*1*). We recently conducted interviews on the state of data-informed decision-making with university leaders who were highly attuned to how well their institutional data systems and organizational structures are serving them and to the kinds of data capture and exploitation most needed. Findings from this exploratory study shed light on ways in which universities are data rich, data poor, and—sometimes—intentionally data blind. They point toward the need for leadership that supports a panoramic view of the data infrastructure and policies at play within individual universities, whether realized by creating a new senior role with relevant authority and budget or through greater multistakeholder coordination.

The cost of poor data management and the lack of data governance is an invisible tax on an organization's efficiency. Despite sporadic initiatives in recent years to grow interoperability and reduce redundancies in academic data management, most institutions still lack needed coordination and expertise. Over the past two decades, a commercial market has emerged for expensive systems that manage information about instruction, scholarship, grants, human resources, finance, and operations. Our engagement with university administrators revealed both concerns about commercial control of their internal systems and continuing tensions about local capacity for data-informed planning. Many felt handicapped by the lack of databases of record, coordinated information management strategies, and administrators with data science training and skills. Faculty, students, and administrators alike have concerns—some legitimate, some not—about who has access to data, decisions that may result, and potential commercial exploitation of their information. So too, universities have been slower than other economic sectors in creating senior positions such as chief data officers to coordinate data quality, strategy, governance, and privacy matters (*2*). Our study sought to identify sources of these tensions along with innovative solutions adopted or under development within the academy. We unexpectedly found a pervasive void of infrastructure thinking and a relatively limited set of data-informed planning successes.

Although our study did not address the COVID-19 pandemic per se, this unprecedented crisis heightened the salience and urgency of many data considerations. The onset of the pandemic found university administrators scrambling to make data-informed decisions about remote access to services such as health care, instruction, and libraries; about security of buildings, laboratories, and technology; and about the effectiveness of various infrastructures. Myriad privacy issues became apparent as administrators sought aggregated or identifiable information about activities on campuses, networks, and systems.

We interviewed a dozen university leaders selected to represent a balance of perspectives on data management, with roles including provost or vice provost; vice president (or vice chancellor) for research (VPR) or institutional research; university librarian; and chief information officer (CIO) or chief technology officer. Several participants had multiple job titles. Although we interviewed these leaders about their current institutional roles, most of them also commented from their perspectives as current or former faculty. The sample was diversified by type of institution, public or private; by gender and ethnicity; and by geography, with respondents from east and west coasts of North America and midwestern US states (see supplementary materials). We conducted interviews by Zoom, which we recorded and transcribed, averaging about 45 minutes in length, from April through August of 2021.

Our interview questions addressed the participant's role in university data, what key business decisions are data-informed, where they lack data for decision-making, which information systems are most important in making critical management decisions, who is responsible for what kinds of data, what are their criteria for outsourcing or insourcing data systems, what integrative views of university data they need, and where sensitivities about data access and use arise on their campuses. These questions led to wide-ranging discussions that addressed many kinds of data, decisions, strategies, and concerns. Because the United States lacks the centralized models for tracking research outputs and academic productivity common in the UK, Europe, and many Latin American countries, our findings will apply differently by region and institutional arrangements. Similarly, comparisons to government and industry are inherently limited. University infrastructures must accommodate a complex array of stakeholders, missions, data resources, and time horizons.

URGENT CHALLENGES

Participants spoke to the urgency of the data governance and exploitation challenges faced by universities. In coding the long lists of data elements mentioned in our interviews, three general categories of data emerged, varying by origin, application, and policy sensitivity (see the box). Various interdependencies arise among these three categories of data, often requiring interoperability between systems. For example, for library collections to support the teaching and research missions of the university effectively, their systems incorporate telemetric and administrative data from internal systems for learning management, registrar, identity management, and finance and may interoperate with external systems of publishers, community repositories, and other agencies.

Data for strategic decisions

The ability to monitor activity related to teaching, research, telecommunications, building services, and operations proved crucial in transitioning to remote work at the height of the COVID-19 pandemic. Crisis experience also revealed where data to inform decisions were lacking. Valuable

[1]Department of Information Studies, University of California, Los Angeles, Los Angeles, CA, USA. [2]The MIT Press, Massachusetts Institute of Technology, Cambridge, MA, USA. Email: amybrand@mit.edu

data may be inaccessible because of data governance practices or friction between stakeholder groups, or be technically accessible but not exploited because of a lack of staff expertise. One VPR told us, "I don't think we made very good use of our data, in part because the people we had in our office were not used to using data to make decisions."

Provosts, librarians, VPRs, deans, and faculty all want better control over records of scholarly products such as publications, research data, materials, and software. In some universities, faculty productivity data are held centrally in "research information management systems" or "current research information systems"; many of these are commercial products (*3*). In other universities, faculty productivity information is decentralized, held by deans and departments.

Provosts told us that they could make more strategic hiring and curricular decisions if they had more comprehensive data on faculty research areas, career interests of prospective students, research funding patterns, higher-education policy trends, and competitive intelligence about other universities. Librarians could target their collections and services more strategically with current data about faculty scholarship and curriculum decisions. VPRs could grow their university's funding and scholarship portfolios more effectively with fuller data on current and emerging research areas, faculty and staff expertise, and tools to match talent to funding opportunities. Faculty members, per our participants, could benefit from tools that reduce their administrative burden to produce data for populating personnel files, institutional repositories, and university reports. In STEM fields, faculty also seek data about potential collaborators.

Most participants sought more transparent and coordinated approaches to their university's data resources. Our overall findings supported the value of chief data officers, although only a few participants explicitly expressed a need for such a leadership role. Arguments for taking an infrastructure approach to university technology investments include more effective use of data for strategic decision-making; decreased duplication of systems and labor; and policy oversight for privacy, data protection, and cybersecurity.

Barriers to access, control, and use

One clear point of unanimity among participants was that tensions exist among stakeholders regarding who has access to particular data, appropriate uses of data, and mechanisms for data governance, privacy, and integration. We highlight key issues here.

Data governance and privacy

In policy contexts, data governance spans privacy, security, data protection, boundaries, ownership, authority, stewardship, and degrees of centralization versus decentralization. Participants offered a variety of opinions about who should control which kinds of data, systems, and uses. One provost commented that "COVID has changed our practices and expectations about control over data, such as Zoom, and who has access to transcripts and recordings." Faculty and other stakeholders are concerned about the surveillance capabilities of centralized systems and how integrated data might be used for resource allocation.

Categories of data

Telemetry data
Functional administrative data—the signals generated by systemic transactions, wireless networks, security cameras, sensors, and similar information produced in day-to-day operations.

Academic administrative data
Information about students, faculty, staff, visitors, collaborators, funding, and academic outputs such as publications. These data span university operations such as personnel, research information management, and learning management systems.

Research data
Products of scholarly activity. These data require local infrastructure and may be subject to sharing requirements by funding agencies and journals.

Decentralization has advantages of local control by individual data stewards who know their data, sources, and users well. Administrators often acquire systems to serve their own needs, with little university-wide oversight for interoperability or data sharing. When local data stewards have sole authority over system access, they can constrain data release to other internal units on grounds of privacy, labor requirements for extracting data, technical incompatibilities, or administrative reasons. We found that data stewards seek to protect their own data: Libraries control patron records, research offices control grant proposals and financial data, provosts control personnel data, and so on. Several participants commented on how decentralized control by strong deans and schools can make coordination across the university "a big problem." One provost said they have many "data czars and czaresses" over their data. Another provost opined that the idea of trying "to have a single personnel system at [our university] … is a fool's errand." Entrenched decentralization constrains an institution's appetite for integration and coordinated planning.

Many governance concerns we encountered were framed as privacy issues, often conflating complex conceptual, technical, and legal aspects of privacy, anonymity, confidentiality, security, and surveillance (*4*). Privacy can be a blunt instrument—or "talisman," as a university librarian said—to assert control over data about individuals. Rarely did we find broad campus engagement in governing privacy matters. Rather, diffused governance of data about individuals was the norm, delegated to offices of the CIO, chief information security officer (CISO), or legal affairs; few universities had chief privacy officers.

Data integration

The difficulty of integrating data across campus was noted by almost every participant. Some integration issues were technical, such as duplicate data in incompatible systems from competing commercial providers. Others were conceptual, such as lack of agreement on data elements, and situations in which too much data integration might impinge on academic freedom. One director of institutional research noted a recent agreement on defining a "gateway course" as a big win. Some integration issues are territorial, such as when one office refuses to share data with another; others were framed as data quality issues, such as inconsistent reporting across units. Administrators need better data rather than more data.

Some integration issues involve claims of intellectual property rights that conflate ownership, licensing, and educational policy. An example given by provosts and librarians was access to course syllabi. Databases of syllabi would give students a broad view of course offerings and the ability to sample readings, enable provosts and deans to identify duplication and gaps in instructional offerings, and strengthen librarians' ability to match collections to curriculum. Whereas individual faculty may claim ownership of syllabi in principle, in practice these syllabi are available in campus learning management systems—but accessible only by currently enrolled students. A change in policy could open syllabi to the university community. Integration barriers often involve governance more than technology.

Of the six job categories in our study, university librarians were most explicit about their goals for integrating data sources across their campuses. These librarians

want to tailor collections and services to their university's academic strategy and faculty research trajectories (*5*) but report that they rarely are granted access to those data. One university librarian commented that "collection management remains an artisanal practice," for reasons such as the ways in which "collection purchasing is atomized into autonomous budgets, making it difficult to grasp the big picture." As a consequence, library collection decisions do not depend on data as much as they could.

CIOs, and individuals in other roles, often said that the desire for data integration was greater than the local expertise and political will to invest accordingly. Data integration was particularly complex for universities implementing multiple, often incompatible, corporate solutions. Rather than exporting and exploiting their data resources, they found themselves being forced to "buy back" their own data from vendors, or devoting weeks or months of staff time to merge data manually. Meanwhile, these same vendors are mining and combining university data and open-access data to offer competitive intelligence services to the academic market.

Although no one expected seamlessly interoperable systems across campus, our participants did seek sufficient systems integration to increase efficiency and reduce redundancy, while not forcing interoperability between incompatible systems or across institutional boundaries.

In response to growing concerns about corporate control of academic platforms and analytics (*6*, *7*), institutions are taking great care in negotiating terms for data ownership, control over data migration and integration, documentation, privacy, security, and risk management. More of the private research universities than public universities in our small sample are contracting for systems to manage faculty profiles, library content, teaching and learning services, grants, and institutional research.

Several participants noted that new interoperability platforms improve the ability to exchange data between commercial systems. Others commented that universities cannot compete with business interests in the marketplace for technical talent, making it difficult to maintain and document open-source software or locally built systems. Research universities with medical centers mentioned data integration as a substantial source of friction. Maintaining separate technology operations for central campuses and medical campuses is expensive but often necessary to comply with regulations on health data.

ILLUSTRATION: SABELSKAYA/ISTOCK.COM AND OLGA_ROM/ISTOCK.COM; ADAPTED BY C. AYCOCK/*SCIENCE*

Research data management

Researchers in all disciplines, and especially those who rely on extramural funding, may be subject to requirements for maintaining, releasing, and sharing scholarly products such as publications, data, software, and documentation (*8*–*10*). Despite more than two decades of data-sharing requirements by funding agencies and journals, none of our participants reported coordinated university approaches to research data management (RDM). Much of the infrastructure required for RDM involves workflows throughout the university. One respondent commented, "You need a vice president of research, … CIO, and … a library that's completely on board" to stake out territory between these three entities and "to convince the faculty we can put their data in the safest place." Our participants generally agreed that "difficult conversations among stakeholders are necessary."

VPRs, librarians, and provosts alike deferred RDM responsibility to principal investigators (PIs), disciplinary repositories, and funding agencies. One university librarian commented that it is "hard to justify the high costs of data preservation and stewardship." One VPR in our sample was deeply concerned about funding agency mandates for maintaining access to data: "Agencies are not realistic in what they are expecting of PIs, especially for short-term projects … Maintaining software necessary to use older data (also) is infeasible."

University librarians, to whom several of the provosts, CIOs, and VPRs also deferred, had nuanced explanations of RDM challenges. They now have enough experience to assess methods and costs. Rather than store datasets locally, for example, one library is indexing university datasets that are deposited elsewhere. Another library is developing "workflow wizards" to automate some data management tasks. A VPR is addressing policy to identify data sources worthy of local stewardship and those sources better diverted to disciplinary repositories. Medical libraries have greater oversight of research datasets than occurs in most domains. These are important steps toward addressing the data-sharing requirements of government agencies in the United States, Europe, and elsewhere. However, full compliance remains a daunting challenge to the universities studied, not least because RDM has the classic economic characteristics of a "commons," subject to competition and free riders (*11*). Shared governance models for research data stewardship remain elusive.

LESSONS LEARNED

Academic leaders have legitimate concerns about economic constraints, lack of data expertise, being locked into commercial

solutions, and creating surveillance states on their campuses. Even when their universities are "data rich," they may also be "data poor" in that they are struggling to exploit data resources to their strategic advantage, or "data blind" in being reluctant to initiate stakeholder discussions necessary to build consensus for data governance. Lessons learned from the experiences of these major research universities can help to inform institutions with fewer resources.

Invest in knowledge infrastructures

To improve data access, intelligence, and integration, universities can make greater investments in knowledge infrastructures (KIs): robust networks of people, artifacts, and institutions that generate, share, and maintain various kinds of knowledge—including, crucially, data (*12*). These infrastructures incorporate systems managed by administrative units and their interactions with other technologies and institutions. By mapping existing infrastructure components and points of interaction, universities can identify duplications of effort, sources of friction, and means to improve information flows throughout the institution.

Local and national investments in KI are a matter of strategic competition [for example, (*11*)]. Research data involve more stakeholders than may be immediately apparent, given the many hands, technologies, and university offices touching them: faculty investigators, student and staff researchers, computer centers, grants administration, finance, libraries, institutional repositories, legal and technology transfer offices, and laboratory and building services. These entities in turn interact with their counterparts at partner institutions and with funding agencies, journals, data and software repositories, publishers, professional societies, and other stakeholders in scholarly communication. Small datasets may reside locally on campus but still require complex infrastructure to maintain data, software, code, and documentation. Large datasets, such as climate models, may require multi-institution infrastructure for stewardship (*9*).

Identity management is an essential component of university KIs. As one provost summarized, "without a common identification system, all of this [data integration work] is futile." Universities assign personal identifiers for various privileges, services, access, and finances. An individual might have multiple university identifiers and multiple external identifiers, such as an ORCID and a social security number. In addition to technical and political challenges, mapping identifiers between systems has high stakes for scholarly attribution and credit.

Invest in data management capacity

New data science programs are producing professionals for all economic sectors while also developing data-management expertise in traditional disciplines. As universities begin to employ more of this new crop of professionals, invest in training existing personnel, and develop new career paths for them, they will gain capacity for managing telemetry, academic administrative, and research data. Another benefit of data-management capacity is the ability to mine these data to yield strategic, policy, social, cultural, and technical insights.

Our interviews surfaced many examples of low-hanging fruit, in which small investments can yield large payoffs in data exploitation. These include interoperability tools that extract data from university databases to produce customized information feeds, portals for self-service access to university databases, indices to internal and external resources, and application-programming interfaces (APIs) to access university data. APIs can support search, download, and data visualization capabilities for myriad systems and purposes, such as institutional research, institutional repositories, course materials, and library content. Other APIs are being deployed at universities to index research data, aggregate newsfeeds, populate academic personnel files, and integrate internal websites with external resources such as grant databases. These innovations often arose from seed grants for pilot projects or prizes for student competitions.

Develop FAIR and just data resources

Rarely are data simply "facts." Data collection of all kinds is based on models, assumptions, and methods, whether explicit or implicit. Thus, data-informed decision-making is only as good as the data on which decisions are based. Scholarly research on sources of data bias, and on ways to address them, offers valuable guidance on paths forward (*13*). Engaging the university community in governing data resources in ways that are transparent is an important step toward institutional goals for equity and justice.

The FAIR principles (findable, accessible, interoperable, reusable) (*14*), although adopted widely for research data, also offer aspirational guidelines for KIs. For data to be findable and interoperable, they need to be documented with consistent metadata. For data to be accessible, they must be searchable and retrievable. To be reusable, data need to be associated with sufficient documentation, methods, software, and other tools for others to use them. Although the FAIR principles may be a high bar for telemetry and academic administrative data, the emphasis on interoperability is key to campus data integration.

CONCLUSIONS

National and international policy for data sharing, management, reuse, privacy, and security are advancing rapidly (*11*, *15*). To remain competitive, universities will partner in these initiatives and respond to new policy requirements. As the COVID-19 pandemic has reminded us, good hygiene begins at home. Similarly, good data hygiene begins on campus. Data-informed decision-making provides opportunities to promote transparent governance; advance fairness and equity for faculty, students, and staff; and save money. We encourage university leaders to embrace more objective and transparent data-informed models for decision-making. ■

REFERENCES AND NOTES

1. K. L. Webber, H. Y. Zheng, Eds., *Big Data on Campus: Data Analytics and Decision Making in Higher Education* (Johns Hopkins Univ. Press, 2020).
2. K. Ilkani, "The role of the chief data officer in higher education" (The Tambellini Group, 2020); https://www.thetambellinigroup.com/the-role-of-the-chief-data-officer-in-higher-education.
3. M. Givens, "Keeping up with… research information management systems" (Association of College & Research Libraries, 2016); https://www.ala.org/acrl/publications/keeping_up_with/rims.
4. C. L. Borgman, *Berkeley Technol. Law J.* **33**, 365 (2018).
5. D. Cooper, C. B. Hill, R. C. Schonfeld, "Aligning the research library to organizational strategy" (Ithaka S+R, 2022); https://doi.org/10.18665/sr.316656.
6. C. Aspesi, A. Brand, *Science* **368**, 574 (2020).
7. A. Posada, G. Chen, *ELPUB* 2018, (2018).
8. C. L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World* (MIT Press, 2015).
9. C. L. Borgman, P. E. Bourne, *Harvard Data Science Review* (2022).
10. National Academies of Sciences, *Reproducibility and Replicability in Science* (National Academies Press, 2019).
11. P. E. Bourne *et al.*, *Science* **377**, 256 (2022).
12. P. N. Edwards, *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* (MIT Press, 2010).
13. C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy* (Crown, 2016).
14. M. D. Wilkinson *et al.*, *Sci. Data* **3**, 160018 (2016).
15. National Academies of Sciences, *Open Scholarship Priorities and Next Steps: Proceedings of a Workshop–In Brief* (National Academies Press, 2022).

ACKNOWLEDGMENTS

We are grateful to our participants for their generous contributions of time and expertise. We thank colleagues who served as pilot testers of our interview protocol, provided advice on research design, and commented on prior drafts of this article, including P. E. Bourne, P. Courant, S. Garfinkel, M. L. Kennedy, C. Lynch, and M. Smith.

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.add2734

10.1126/science.add2734

Data blind: Universities lag in capturing and exploiting data

Christine L. BorgmanAmy Brand

Science, 378 (6626), • DOI: 10.1126/science.add2734

Science (ISSN 1095-9203) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science* is a registered trademark of AAAS.