

Libraries and Institutional Data Analytics: Challenges and Opportunities

H. Austin Booth
Vice Provost for University Libraries
University at Buffalo Libraries
433 Capen Hall
Buffalo, New York 14260
habooth@buffalo.edu

Dean Hendrix
Associate University Librarian for Research, Education and Outreach
University at Buffalo Libraries
232 Lockwood Library
Buffalo, New York 14260
dhendrix@buffalo.edu

The archived file is not the final published version of the article.

The final published version is: Booth, H. Austin and Dean Hendrix. "Libraries and Institutional Data Analytics: Challenges and Opportunities." *The Journal of Academic Librarianship* 41, no.5 (2015), 695-699.
[doi:10.1016/j.acalib.2015.08.001](https://doi.org/10.1016/j.acalib.2015.08.001)

Libraries and Institutional Data Analytics: Challenges and Opportunities

Libraries are well positioned to lead, facilitate, manage, or play a major partnership role in institutional data analytics programs. The degree to which an institution can create a successful data analytics program, and to which libraries can lead such a program, is highly dependent on institutional culture and the role the libraries play in strategic on-campus partnerships. What follows are the challenges and opportunities for libraries helping to create robust institutional analytics programs.

Why analytics?

When our institutional colleagues hear the word “analytics,” other words that come to mind are “data,” “analysis,” and “decisions.” Libraries can help to create a culture shift so that “analytics” also brings to mind “strategy” and “results.” Analytics is not just data, it is a process to use data to answer strategic questions and make strategic decisions in order to produce strategic results.

The best uses of analytics are those that let you ask and answer:

- what will happen if trends continue (forecasting);
- what will happen next if you take a particular action (predictive modeling);
- what is the best solution to a particular problem (optimization).

Greater calls for accountability in higher education have led to an increased use of data analytics on our campuses to document the impact of higher education-- impact on students, local community, academic or professional disciplines, and ultimately, larger social concerns such as economic development and health. Another driver for using analytics in higher education is the growth of the use of big data and continuous improvement processes in almost every other industrial sector.

Currently, higher education institutions are primarily using data analytics in four ways: to optimize student success; to manage enrollment; to optimize resources (to improve the performance, efficiency, effectiveness of the institution, particularly on the academic support or administrative side); and to understand and grow the research enterprise at their institutions.

In all four areas, analytics are increasingly being used to allocate resources, which means the stakes are high. Colleges and universities are using research productivity metrics, for example, to communicate research quality, impact and value to secure funding, receive tenure and promotion, earn incentives, increase salaries, recruit high caliber graduate students, and grow their laboratories or research centers. Academics unable to show their effectiveness using quantitative and qualitative data may have a hard time obtaining employment or securing resources.

Challenges:

Even though higher education institutions are collecting more data than ever before, much of this data is data required for mandatory and/or external reporting, not data that could be useful in answering institutional strategic questions. Furthermore, much of the data that is collected is not used at all. In other words, if you think that your institution is behind in its attempts to use analytics, you are not alone (Bichsel, 2012).

The challenges to creating a robust institutional data analytics program include culture, talent, cost, and data. We have deliberately mentioned culture first because it is very easy to jump to data challenges. In fact, most of the literature surrounding data analytics starts with challenges surrounding the data itself. However, we are convinced that institutional culture is the most important factor in determining the success of any given data analytics program, including the politics and process around questions of talent, cost, and data itself.

Culture:

Data-driven cultures are not a given, and creating a data-driven culture is not easy.

There is, on many campuses, a widespread mistrust of institutional data, measurement, analysis, reporting, and change (and the conflation thereof) across administrators, faculty, staff alike. An article in *Nature* reported that 63% of researchers and administrators expressed unhappiness with the use of metrics in higher education (Abbott, et al., 2010). The mistrust ranges from a feeling that higher education is not a good business—that it does not know how to use data to make decisions, to the opposite—that higher education is becoming more of a business, that analytics is the ultimate sign of that change, and that this change is counter to the values of higher education.

Mistrust is especially acute when discussing research performance. When stories circulate in the academy about researchers or journals gaming citation data or the Matthew Effect of well-known journals, faculty's perception that these metrics are invalid and promote corruption is reinforced. Furthermore, many faculty believe this scorekeeping mentality encourages unhealthy competition and a homogenization of "safe" research at their institutions in lieu of collaboration and novel experimentation. Finally, there is an understandable resentment of having careers represented by narrow measures, as may be the case for institutions using journal impact factors or the h-index. Despite these objections, mistrust in analytics can be mitigated if faculty are given meaningful roles in defining the metrics (Wilsdon, 2015).

There is a growing body of fascinating and highly valuable research on how libraries can link their work to the institutional data that institutions are focusing on (e.g., mapping library instruction to student success metrics); here what librarians are interested in is the broader context of data analytics on our campuses and the factors that are important to consider in helping to create successful and useful data analytics programs on our campuses. Again, libraries can play a major role in the creation of such programs, beyond and including data about libraries themselves.

Another challenge is the “Analytics IQ” of leaders in higher education, from top administrators down to supervisors in units, is not high. Institutions tend:

- not to recognize the degree to which the use of data analytics—particularly performance-based analytics or efficiency-improvement analytics—requires a significant culture change which will need to be managed using culture-change management techniques;
- not to understand the gap between the institutional capacity for data and analysis and what will be required to undertake large and meaningful analytics projects, particularly in terms of the expertise needed for data analysis, and the simple amount of time and human—rather than machine—labor it takes to do quality analytics work; and
- not to be realistic about expected outcomes;

Finally, data governance is a key part of the culture surrounding data on any particular campus. By data governance, we mean “the organizing framework for aligning strategy, defining objectives, and establishing policies for institutional information” (Dyche & Nevala, 2014).

The challenge is that many institutions have not defined data governance and confuse it with data management. Data governance is the designation of decision-rights and policy-making surrounding institutional data, while data management is the implementation of those decisions and policies. Institutions need both, and both require investment, but the senior leadership of our institution needs to design the former.

There is no “best practice” for data governance—your data governance needs to reflect and support your institutions’ values and behaviors; decision-making structures (both explicit and implicit, formal and informal); design and implementation processes; tracking and assessment policies; ownership culture; collaborative history; communication practices; work rhythms—in other words, its ways of doing business: its culture.

Talent:

A big challenge to developing this culture is that, across all sectors of the global economy, there is a lack of skilled big data professionals (Manyika, 2011). It is a big challenge, not only that there is a low supply of those with the proper knowledge and expertise, but that the need for that talent is underestimated. Librarians need to create programs within our institutions or look outside for this talent, and they need to work with other organizations, such as library and information science schools, to ensure that we have a pipeline for these professionals in the future.

Any data analytics work occurring on our campuses resides in institutional analysis offices, IT, extramural research offices, student services, registrars’ offices and in the academic departments. It is important that all of these groups participate on a central data governance

council to develop and mold a central data strategy. For instance, institutional analysis office staff are often consumed by external mandatory reporting and therefore simply do not have the time and energy to undertake additional data analytics projects. The use of data generated for external reporting purposes is not generally the same data that would be useful to address strategic institutional questions because of data definitions collected (which generally aggregated at too high a level). We must think not only about hiring people with these skills, then, but about where they are going to reside institutionally—that is, who determines what their priorities are going to be. The priority setting of data analytics work, is, of course, a cultural question.

It is also important when arguing how analytics are going to be used to streamline procedures, and create efficiencies that will result in decreased workloads (and the ability to reallocate resources, that we, as well as our senior leaders, pay attention to the workload involved in doing data analytics projects. Because senior leaders do frequently underestimate the expertise and time required to undertake meaningful data analytics projects. If librarians take on new tasks, it is very important to track the amount of time and level of staff required when undertaking analytics projects. For example, collecting citation data for a researcher with a common name often requires manual and painstaking record- by-record searching in order to disambiguate that individual's research from others that share his/her name. This type of work requires a librarian with a deep and intimate knowledge of the bibliometric databases that are being used to harvest the bibliometric data.

Cost:

Data analytics should be thought of as a strategic investment, not a cost-saving technique. When we talk about barriers to creating effective analytics programs and we talk about cost, usually we are usually talking about the cost of tools and data collection methods, when the real cost is the time spend on cultural change and on developing and educating a staff with the analytical skills we need in our discipline. Moreover, a visionary analytics plan invests in people, in hiring and training, over data tools and platforms.

Again, the real challenge is that we need senior leaders to see analytics as an essential element of strategic planning; it is very important to argue that analytics can help senior leaders achieve strategic institutional goals rather than simply achieve cost savings.

Data:

Much energy and time is spent on our campuses engaging in data debates rather than in formulating and answering strategic questions or addressing issues that might be solved with the use of data. This often leads to an overreliance on locally collected and synthesized data. This may be done to distribute workloads evenly and assuage political concerns, but with departments or units inside each institution using their own data definitions and processes to collect and transform data, analytics are disorganized, irreproducible, ineffective, and at worse, meaningless. Repeat: Data wars are a culture problem, not a data problem.

The four data challenges that we face most often when creating/discussing institutional data analytics programs are:

- ownership
- quality,
- standardization, and
- access.

Ownership

In order to run a successful analytics program, institutional data needs to be taken out of its silos and centralized. Many units, especially those experiencing existential angst, are hesitant to share data for fear of losing control of the data and more importantly, hypothetical repercussions. Clear use statements from senior leaders regarding how data will be used can help in creating a greater readiness to share data. Again, although there are certainly technical issues involved in sharing and merging institutional data, ownership of data has more to do with institutional culture than anything else.

Quality

Units across our campuses are concerned about data accuracy and integrity. Because data wars can be significant barriers to the creation of successful data analytics projects, a process must be established for cleaning the data, prioritizing which data elements are most important to be “accurate,” and deciding when data is “good enough” to proceed. Strong data governance is key here. For example, many faculty refer to their Google Scholar citation counts as their “official” count as it often exceeds those found in Web of Science or Scopus. However, librarians understand the quality control issues when relying upon Google Scholar for accuracy. They also understand that Google Scholar data may also add distinctive elements to defining research impact. Data does not need to be perfect to create a successful data analytics program. Great care in building a culture, which values careful design and deliberate implementation is the most important action an institution can take to ensure data quality.

Standardization

Developing consistent definitions for data elements and the creation of data dictionaries will help to build confidence in data analytics process and encourage (and ease) the sharing of data. Creating consistent definitions is highly complex. For instance, in the area of research metrics, disciplinary data definitions must be addressed especially if comparisons are being made across an institution. In these cases, standardization must be addressed in a consistent collection process and transformation of data, or relational data dictionaries must be created.

Access

In a successful analytics program, data needs to be centrally accessible. The notion of data accessibility, again, goes back to institutional culture: to an articulation of principles of transparency and policies around data roles and responsibilities. A central data repository should be a key action item on any data governance council’s agenda. The creation of a single

data repository that is both transparent and accessible does not mean that units will not maintain local data, but local data can be easily linked to a central repository. Even if a central data repository cannot be created, mechanisms for data transparency and access for local/unit-based data can be created.

Many of us work on campuses where so-called “shadow databases” proliferate. In fact, libraries maintain our own “shadow databases.” We prefer to call these “local databases.” We maintain these for good reasons: we produce and use data that the larger institution is not interested in (or we assume it is not), we cannot get the data we need from our central institutional analysis offices (or we cannot get it in a timely fashion), or we need a different version of the data that our central institutional offices produce (e.g., faculty counts in the “biomedical sciences” that we use for licensing purposes). Due to a variety of reasons, ranging from securing grant funding to professional narcissism, many faculty collect and synthesize their own personal research performance indicators locally which can create confusion when actionable metrics are needed.

Opportunities for libraries:

The growing interest in higher education analytics presents an opportunity for libraries and librarians to proactively address the culture, talent, cost and data issues in institutional data analytics programs.

Leadership in the analytics realm is a natural fit for academic libraries for a myriad of reasons:

- Libraries already collaborate with both the academic and academic support side of the university;
- Libraries have critical partnerships with many of the commercial vendors who are producing the analytical tools needed for this work;
- Libraries know the politics and policy issues of working with commercial parties regarding our own research and research data;
- Libraries manage data and understand data lifecycles;
- Libraries have a user-centered culture, know campus user groups, and know how to design programs around their needs;
- Libraries know metrics, including the challenges and benefits of their use;
- Libraries have historically needed to prove their value;
- Libraries know metadata;
- Libraries know assessment;
- Libraries know strategy;
- Libraries are neutral;
- Libraries are involved with both the research and the teaching and learning sides of our colleges and universities;
- Libraries can bring particular expertise in faculty productivity metrics to the table;
- Libraries have influence on our campuses; and

- Libraries are part of larger professional bodies (e.g., ARL, ALA) that can influence the shape of data analytics.

On many campuses, assessing the research enterprise offers libraries and librarians the most natural entry point to involve themselves in institutional data and analytics discussions due to our expertise in bibliometrics, bibliometric databases and organization of scholarly knowledge.

Culture

The two keys to building the culture that will enable a successful analytics program to function are leadership and governance structures.

Leadership

The creation of a data-driven culture needs to start at the top. Senior leaders need to drive the data analytics program. The wide-ranging relationships that librarians have across campus allow for them to lead these efforts and influence other decision-makers.

In order to create a successful data-driven culture, libraries must identify those senior leaders who:

- have decision making authority and the power to demand accountability; around the implementation of those decisions;
- experiment with new tools, solutions, and practices;
- have been successful in leading culture change efforts;
- begin with strategic questions before jumping into data collection;
- have a track record of relying on evidence to make decisions rather than anecdote, tradition, hearsay, “experts,” previous experience, or intuition; and
- use a performance review/assessment system.

Governance:

Libraries can play a large role in creating a data governance structure and play a leadership role within that structure. Without clear data governance structures, data analytics projects become mired in data wars, debates over roles and responsibilities, confusion over goals and priorities, ad hoc projects, and unanticipated backlash to culture change.

Not only does the data governance structure need to assign clear responsibilities, but it needs to establish a clear prioritization process so that data implementation groups are not lost in a seemingly infinite list of questions and issues that may (or may not) be addressed through data analytics techniques. In other words, a data-driven culture that is based in the identification of strategic questions and priorities that establishes clear decision-making protocols must be developed upfront in order to be successful.

Talent and Cost

We need to develop a range of professionals who can assist with transforming strategic questions into data analysis projects and understand underlying policy and process issues as well as those who can implement projects and develop appropriate analytics deliverables, such as reports, alerts and dashboards. Rather than relying on institutional analysis office staff, a data governance structure can allow for a more distributed use of analytics talent that may already exist on your campus. Institutional analysis office staff may be able to develop dashboards that would enable designated campus unit staff to run their own analytics reports rather than having to rely on central office staff to run ad hoc reports as requested. For example, librarians with bibliometric expertise can provide campus-wide education on tools and unambiguous processes for data collection and analysis.

The data governance group may also assign responsibility for relatively simple data requests such as enrollment counts to another campus unit (central information technology units or academic support units). Minimally, the data governance group should establish a way of prioritizing the requests made of offices of institutional analysis.

Another use of distributed staff is to share programming duties associated with academic analytics efforts—experts in units or even student programmers may be employed for programming expertise. If a unit wishes to create analytics, the office of institutional research may be used as an advisor, guiding the creation of analytical tools that will be used at a unit or central level, rather than as the creator of those tools or the supplier of the sought-after information.

The institutional analysis office or the data governance group may also be used to bring together a community of experts from across campus who are experts in different types of institutional data (e.g., student performance, enrollment, space, research productivity, etc.) to share locally-created tools and practices.

We need to invest in staff and the development of data policies and procedures rather than focusing our investment on new data analytics tools or simply more data collection efforts. We also need to frame the conversation around the expense of analytics as an investment rather than a cost.

Data

As part of the data governance process, responsibilities around the following data functions should be assigned and librarians have much to offer. These include:

- Creating an inventory of institutional data;
- Developing a data dictionary;
- Designing an unambiguous process for cleaning up those data elements designated as most important to answering the strategic, prioritized questions determined by senior leadership; and

- Creating an open data set that answers to the most commonly asked data questions across campus.

Ownership

In the highly charged environment of institutional analytics, libraries are oftentimes seen as a neutral party, and can educate administrators on the need for data governance and data management at the institutional level. This may help also in asking distributed departments to actively participate and share data.

Quality

For example, librarians can easily lead campus conversations on research data quality and integrity as well as assist with the preparation of bibliometric reports for individual faculty or departments. Additionally, providing bibliometric education services is another critical way libraries can serve their campus communities. Education may include offerings on institutional standards for the collection of data, evaluative analysis of tools and performance metrics, and the emergent role of altmetrics in some disciplines.

Standardization

Publication-based research performance indicators, such as citations, the Journal Impact Factor (JIF), and the h-index, are now a part of many disciplines' research lifecycle. Academic librarians have been active partners in the creation of standards for useful and actionable research performance indicators. The most prominent example is the Snowball Metrics initiative that provides data definitions and "recipes" that are reproducible across institutions. Of course, problems arise when departments use different data sources (e.g., Web of Science versus Scopus versus CrossRef versus Google Scholar) to calculate unambiguous metrics such as the h-index. Librarians should also be leaders in developing procedures when data is transformed or synthesized as they understand the downstream consequences of short-sighted ad hoc processes.

As one of their core responsibilities, librarians should be familiar with the tools and the underlying data behind each tool and provide thoughtful evaluations of these information products in the analytics space. Libraries can also coordinate and implement unique, verified and unambiguous author identification standards to insure research data quality, comprehensiveness and utility. Several libraries across the country, including the University of Michigan Libraries, are leading the ORCID implementation on their campuses (University of Michigan Libraries, 2015).

Access

Just as they do with libraries collections, librarians can help their parent institutions in making data more easily discoverable by organizing and describing data collections. Currently, if done

at all, the assignment of metadata to institutional data is uneven and has no syndetic hierarchy underlying it.

Conclusion

Before undertaking the creation of a data analytics program or even creating a data governance council, assess your institution's readiness for analytics by asking:

- where is your institution in terms of analytics;
- what are the barriers/challenges on your campus; and
- where do you have capacity for developing or expanding current analytics programs?

The higher education consulting firm, the Voorhees Group, has an extensive institutional data readiness assessment instrument that can help in this process:

<http://www.voorheesgroup.org/tools/institutional-data-readiness-assessment-tool/>

Librarians have often undervalued themselves in the knowledge management and institutional data analytics space. Applying our professional expertise in managing information resources and building processes to make data and metadata uniform, discoverable and open put us at the fore of these conversations. The political nature of this work cannot be underestimated. It is highly charged and those strong feelings across your institutions should signal that it is important work – work that we are experts in.

References

Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q., & Van Noorden, R. (2010). Metrics: Do metrics matter?. *Nature*, 465(7300), 860-862. DOI 10.1038/465860a

Bichsel, J. (2012, August). *Analytics in Higher Education: Benefits, Barriers, Progress, and Recommendations*. Louisville, CO: EDUCAUSE Center for Applied Research. Retrieved from <http://net.educause.edu/ir/library/pdf/ers1207/ers1207.pdf>

Dyché J. & Nevala K. (2014) *Ten Mistakes to Avoid When Launching Your Data Governance Program*. *SAS Best Practices White Paper*. Retrieved from http://www.sas.com/en_gb/whitepapers/ten-mistakes-to-avoid-when-launching-data-governance-program-106649.html

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A. (2011, May). *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. Retrieved from http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

University of Michigan Libraries. (2015). ORCID Initiative. Retrieved from <http://www.lib.umich.edu/orcid>

The Voorhees Group. *Institutional Data Readiness Assessment Tool*. Retrieved from:
[http://www.voorheesgroup.org/voorheesgroup-
tools/Institutional%20Data%20Readiness%20Assessment%20Tool.pdf](http://www.voorheesgroup.org/voorheesgroup-tools/Institutional%20Data%20Readiness%20Assessment%20Tool.pdf)

Wilsdon, J., et al. (2015, July). *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. DOI: 10.13140/RG.2.1.4929.1363 –
Retrieved from
<http://www.hefce.ac.uk/pubs/rereports/Year/2015/metrictide/Title,104463,en.html>