SPARC

A Roadmap for Action

Academic Community Control of Data Infrastructure

November 2019

Table of Contents

ACKNOWLEDGEMENTS

SPARC would like to acknowledge that the development of this publication was generously supported by grants from the Open Society Foundations and Arcadia – a charitable fund of Lisbet Rausing and Peter Baldwin. We are also deeply grateful to the following colleagues for sharing their expertise and comments during our process of developing this document:

Amy Brand
Chris Bourg
Marilyn Billings
Federica Cappelluti
Boyoung Chae
Leslie Chan
Tom Cramer
Mercé Crosas
Spencer Ellis
Ellen Finnie
Elena Giglia
Jean-Claude Guédon
Joy Kirchner
Elizabeth Kirk
Rick McCollough
Salvatore Mele
Roger Schonfeld
Peter Suber
Martha Whitehead

BACKGROUND

Earlier this year, SPARC released an in-depth Landscape Analysis of the changing academic publishing industry and the implications of the large-scale deployment of data and data analytics. The the academic community's response was immediate and strong, and it underscored the need for coordinated and strategic action to avoid the potential consequences. As noted in the Landscape Analysis:

> "Until now, [commercial publishers] were – at worst – seen by institutions as an annoyance for selected communities within academia. Librarians complained about the cost of periodicals and talked about a "serials crisis", but the impact on the overall budget of a university was well below half of a percentage point. Similarly, the high cost of textbooks was an issue for students, and in particular those coming from disadvantaged backgrounds, but scholarships and some forms of financial aid, as well as the used textbook market, tended to mitigate the problem.
>
> The move by publishers into the core research and teaching missions of colleges and universities, with tools aimed at evaluating productivity and performance, means that the academic community could lose control over vast areas of its core activities. In addition, the collection of massive amounts of data about faculty and students poses a significant legal and reputational risk for institutions, along with potential privacy and security threats for individuals."

As the project unfolded, we recognized the need for a range of potential solutions for the key stakeholders to consider pursuing. The purpose of this document is to build on the Landscape Analysis[1] by offering a roadmap of potential actions that stakeholders can use to chart both individual and collective responses. Recognizing that solutions to these complex issues are not "one size fits all," this document offers a framework with multiple, concrete solutions that individual organizations can improve and adapt to their local culture and needs. The solution set is by no means exhaustive, and is intended as a starting point for the community to build upon.

[1] https://sparcopen.org/our-work/landscape-analysis/

We are pleased that this project has helped to stoke a debate around the need for academic institutions to take concrete steps, and that our initial analysis has proven to be a valuable resource on campus. As an example, this spring, as the University of California system grappled with the issues surrounding adopting commercially-owned research information management systems, its Academic Senate issued a set of recommendations which refer to and largely support the initial findings and recommendations contained in our Landscape Analysis.[2] We hope that the academic community will find this document a useful additional resource to inform discussion of these issues on campus and to serve as a catalyst for taking considered action.

Report Authors:
Claudio Aspesi (Lead Author)
Nicole Allen
Raym Crow
Shawn Daugherty
Heather Joseph
Joseph McArthur
Nick Shockey

[2] https://senate.universityofcalifornia.edu/_files/reports/rm-jn-mb-rims.pdf

WHAT DO WE MEAN BY DATA AND DATA INFRASTRUCTURE?

Research Data and "Grey Data"

In this document, we talk about two types of data. The first is Research Data, which refers to the data academic institutions generate through their research activities. The second is Grey Data, which refers to the vast amount of data produced by universities outside of core research activities, and which tends to focus on the individuals belonging to its community (primarily students – but also faculty and staff). This includes data from applications, student records, ID cards, surveys, sensors, surveillance video, internet and network usage, etc. The term "Grey Data" was coined by Dr. Christine Borgman[3] at the University of California Los Angeles, who points out that the boundaries between research and Grey Data are increasingly blurry, and it is necessary to consider both when discussing solutions to the issues posed by the rise of data and data analytics in academic institutions. As a result, this document addresses both types of data.

It is also critical to underscore upfront that we are not opposing the use of data and data analytics in academic institutions. Data collection and analysis are key elements of research, teaching, and learning.[4] We do acknowledge that there are voices arguing that the unbalanced use of machine learning can shift the focus within the academic community away from basic science and into technology - or away from theory and into producing large data sets.[5] However, a world without data is also a world where biases can and do play a large role, so limiting the use of data is not the solution.

[3] https://arxiv.org/ftp/arxiv/papers/1802/1802.02953.pdf

[4] SPARC staunchly advocates for Open Data as a way to accelerate the research process, but our advocacy for Open Data policies in independent of our work in this document.

[5] https://medium.com/berkman-klein-center/from-technical-debt-to-intellectual-debt-in-ai-e05ac56a502c

Our goal is to ensure that academic institutions retain control over the use of their data and of the analytics applied to it. It is also vital that their use is consistent with the goals of the academic community and that academic institutions are properly equipped to deal with the risks and implications posed by the use of data.

Of course, in many ways, this phenomenon mirrors the rise of data capture and usage in society, and it poses similar challenges. What is different is the declining opportunity for individuals within academic institutions to actively opt out of data collection. Individuals working for corporations, depending on where they live, have limited or no expectations about digital privacy at work. On the other hand, academic institutions, at least in many western countries, have always protected academic freedom, including the right to conduct research and search for information, without prying eyes. These concepts are now at risk.

Metrics and algorithms

It is important to distinguish upfront between metrics, which refers to what is being measured, and algorithms, which refers to how it is being measured. These two categories are often interrelated, but they pose different issues, and therefore, should be addressed separately.

Metrics should be controlled by academic institutions. It is their responsibility – and theirs alone – to ensure that evaluation is performed on the basis of multiple factors that align with the institution's mission and values. This document does not advocate for academic institutions to choose any specific metric, but rather simply that they deliberately choose what metrics are used, rather than simply relying on those sold by commercial vendors. Developing metrics may be complex, resource-intensive, and unique to each institution's context. However, the sharing of best practices across like-minded institutions may facilitate the establishment of a number of metrics which could become de facto standards.

Algorithms, on the other hand, do not necessarily need to be controlled by each academic institution, but must be carefully understood and monitored. It is critical that algorithms are as transparent as possible, so that they can be fully analyzed and held accountable.

So long as an algorithm remains a “black box,” an institution is powerless to understand whether it may contain biases that are incompatible with its values, or flaws that could lead to costly mistakes.

THREE CATEGORIES OF ACTION

Any proposed solutions must address the different challenges posed by the adoption of data analytics in research, teaching, and student life. The menu of proposed solutions is based on two complementary organizing principles. The first, already delineated in the previous section, distinguishes between metrics and algorithms. The second organizing principle distinguishes between three types of action, based on the level of control that academic institutions have. We have called the three categories respectively risk mitigation, strategic choices, and community actions. Each category is summarized below, then discussed in more detail in the remainder of the document.

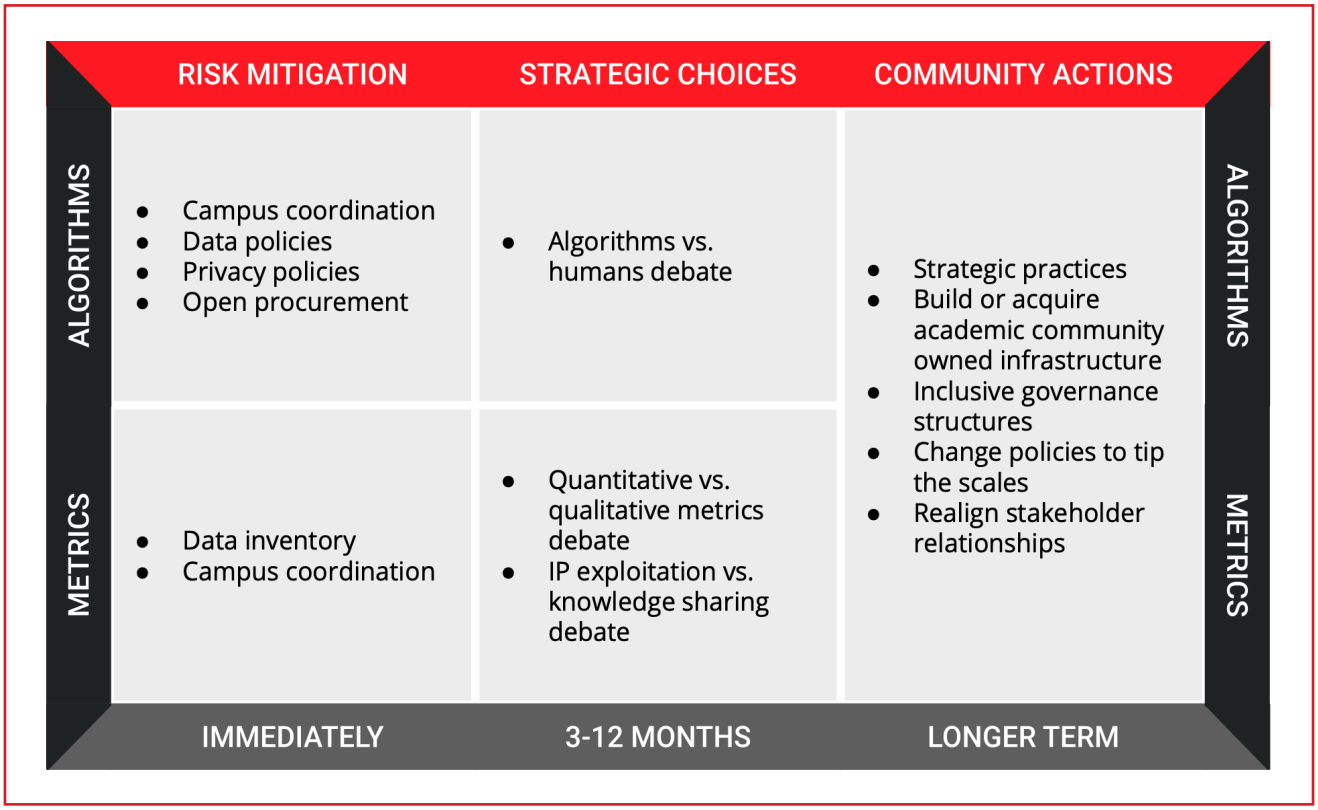
1. **Risk Mitigation.** Under risk mitigation actions, we include steps that individual academic institutions can take to mitigate, if not eliminate altogether, some of the risks posed by the collection of data and deployment of data analytics. These actions often can be executed relatively quickly, require a varying but manageable amount of investment and expense, and achieve a tangible impact on how data is treated within institutions. Examples of these actions include the establishment of coordination mechanisms on campus, revision of data policies, and the adoption of open procurement policies.

2. **Strategic Choices.** Actions in this category require a thorough debate of issues that do not have easy answers. Questions including what metrics to use, the extent to rely on artificial intelligence, and the extent to which to prioritize IP exploitation require a much deeper analysis of both pros and cons, as well as a realistic assessment of what is culturally acceptable in terms of actions. Individual Institutions will legitimately have very different responses to these choices according to their values and missions, and will need to engage a wide variety of stakeholders. These debates are often centered around cost-benefit analyses of widely adopting data analytics tool.

3. **Community Actions.** Community-based, structural actions are possible end-game solutions that may allow groups of institutions to retake control of their data infrastructure. This category includes different scenarios, largely depending on the resources available. There are possible trade-offs between speed of rollout (which could be achieved by acquiring existing infrastructure) and amount of investment (since building the infrastructure from the ground up may prove more cost effective). Also included in this category would be the pooling of intellectual property and text and data mining of research to develop insights that are valuable to industry and financial institutions, or negotiating collective deals with better terms, pursuing policy actions, and realigning stakeholder relationships.

These categories are directional, rather than prescriptive, as we actively support the availability of a wide array of infrastructures. These can be built from scratch by the academic community, can be acquired and grown, and they can also be managed by commercial vendors and/or funding bodies willing to work with the academic community on innovative joint governance models. In turn different institutions will make different choices on which initiatives to support, how, and when.

Open Data Infrastructure: A Roadmap



1. RISK MITIGATION

Risk mitigation actions can be described as those which individual academic institutions might take as a common-sense response to the increasing volumes of data collected across campuses and the rising deployment of data analytics tools. These actions are designed to be concrete, practical steps that any institution can begin taking immediately. Differences among academic institutions may lead to adopting different solutions, and this section lays out a variety of options to consider.

Conduct a Data Inventory

An important step for any institution to consider is to conduct an inventory of what data is collected across campus, how it is collected, and where it is housed. This also includes an explicit accounting of which offices or departments have agreements or contracts with third parties that involve data. As research and grey data become an increasingly integral part of how institutions operate, there is a great risk of losing control of this data if not tracked and managed across the often distributed operating architecture of higher education institutions. Many institutions have not yet conducted such an inventory, and the task may seem onerous at first. However, an accurate inventory is a prerequisite for next steps.

While individual institutions may have varying approaches for conducting such inventories, the process has several commonalities. It generally begins by designating a person or team to be responsible for the following activities:

1. Conducting and maintaining an inventory with standard information on what data is collected, by whom, in what formats, and for what purposes.

2. Evaluating the quality of the data, including how it is generated, how often it is updated, any issues that arise with its generation and aggregation, and its consistency across departments/offices/schools/campuses.

3. Developing plans to standardize data collected across the institution when needed.

4. Coordinating with the procurement and legal departments to analyze contracts or agreements with third parties (regardless of whether they are commercial vendors, other academic institutions, funding bodies, or governments) to understand the contractual obligations of each party connected with data and data analytics on campus.

The successful completion of such an inventory lays a strong foundation for subsequent risk mitigation actions to be taken.

Establish Campus Coordination Mechanisms

The establishment of a coordination mechanism to adjudicate conflicts among departments and offices on data and data analytics contracts is an important element. Coordination is a relatively easy task to articulate, but a complex one to execute given the decentralized culture of many academic institutions. In a more centralized corporate environment, the solution is increasingly to designate a Chief Data Officer (CDO),[6] which is typically defined as a senior leadership role reporting directly to either the CEO or COO of a corporation. This model is also being adopted by the U.S. Federal Government: the Open Government Data Act passed in 2018 explicitly requires that each federal agency designates a CDO, and that the White House Office of Management and Budget (OMB) establishes a Chief Data Officer Council (which includes the CDOs of each individual agency as well as some appointees of the Director of the OMB). [7]

As a preliminary step, institutions may consider establishing a temporary coordination task force or committee as a stepping stone to habituate the institution to the need for coordination on data issues. This coordinating body might include the people responsible for data decisions within each office or department that generates, acquires, externally releases, or stores significant quantities of data, as well as the CIO or CTO of the institution, the Chief Legal Counsel, the individual in charge of strategic

[6] https://www.gartner.com/smarterwithgartner/3-top-take-aways-from-the-gartner-chief-data-officer-survey/

[7] https://www.congress.gov/bill/115th-congress/house-bill/4174/text#toc-H8E449FBAEFA34E45A6F1F20EFB13ED95

planning, and representatives of the library. Once an institution takes the next step of appointing a CDO, this task force could continue acting as a support or stakeholder liaison group. The regular involvement of Presidents or Provosts is also important, although they do not necessarily need to be members of the task force.

An example of where a coordinating body would be important is where different interests come into conflict. An office tasked with establishing corporate partnerships may wish to share information and data on early stage research activities of some departments with parties outside of the institution in order to facilitate corporate partnerships. At the same time, the licensing office may wish to keep the same data exclusive to the institution until patents are granted. Both goals are legitimate, but the adjudication mechanism should be a deliberate institutional decision, rather than unilateral departmental choices.

Revise Data Policies

The revision of existing institutional data policies is an important step. Even a cursory review of existing data policies across the higher education institution landscape reveals that many share the same characteristics. They tend to be technical and tactical in nature and define in great detail how to protect different types of data, primarily on the basis of their sensitivity, from unauthorized access. Most will cover student data, but typically only with respect to compliance with federal or state laws. Some extend to intellectual property (IP) generated through research activities, largely to ensure that future claims around the value of the IP can be defended. However, none of the policies we have seen specifically address the strategic issue of regulating authorized access to and use of institutional data.

In short, existing policies tend to focus on preventing unauthorized access to data, rather than ensure that authorized access to data is coherent with the strategic goals of the institution. It is critical for data policies to be revised to address the myriad strategic questions raised by the proliferation of data and data analytics. These questions include:

- What problems/opportunities is data expected to address?
- What data can be shared with different categories of third parties?
- Who should maintain ownership of the data itself?

- What rights should the institution secure with regard to the output of data analysis?
- What resale uses should be allowed?
- What should be the economic goals of data agreements?
- What rights of audit should institutions demand from third parties to ensure adherence to contractual obligations regarding data and data analytics?
- Should open source software be mandated or preferred in order to facilitate transparency?
- Should algorithms be transparent to the institution, allowing users to understand their mechanisms (and possible biases)?

Revise Privacy Policies

The development of strong privacy policies is critical, and must extend beyond legal compliance. The expansion of grey data on campuses has created privacy questions that lawmakers have only begun to grapple with. Particularly in the U.S., the limited legal framework for data privacy leaves it mostly up to institutions to protect themselves and their stakeholders. To ensure that privacy policies address the needs of the community, they should be developed in consultation with the constituencies that will be affected (including but not limited to faculty, researchers, other staff, students, and administration).

There are several data privacy policy frameworks that can serve as a starting point. EDUCAUSE and NACUBO have made available a number of resources to academic institutions to structure their data privacy policies.[8] In addition, Institutional Review Boards (IRBs) may be able to offer useful policies and practices that have been developed or adapted locally.

Some of the key strategic themes that can be covered in strong privacy policies include:

[8] https://www.educause.edu/focus-areas-and-initiatives/policy-and-security/cybersecurity-program/resources/information-security-guide; https://library.educause.edu/resources/2017/5/7-things-you-should-know-about-how-learning-data-impacts-privacy; https://www.nacubo.org/Topics/Privacy-and-Data-Security/Privacy-Data-Security-Resources

1. Banning any unauthorized release of any data on research activities to any third parties, including the government (in the absence of a court order).

2. Requiring any third party which receives or develops student or faculty data to obtain approval from the institution before entering into any agreement to resell or license the data (even in anonymized form), as well as notify the institution of any database breach or government request to obtain the data (with or without a court order). Faculty and students, in turn, should be notified by the institution of any of the above events.

3. Establishing a requirement to obtain student approval to maintain, beyond a reasonable period of time, data gleaned through the use of digital courseware and other services, including time and location of access, patterns of usage, and the learning profile of the students. Student approval should be necessary for the use of any data other than in the course in which the information was collected (for example, preventing student learning profiles to be transferred from one course to the next). U.S. institutions must ensure these requirements are updated to reflect all student data that may be collected, not just that narrowly defined by FERPA.

4. Ensuring students are adequately informed of the possible uses of data collected through digital courseware, access cards, library records, etc. before using these services. This is especially important where the use of a specific product, system, or tool is required for a student's coursework or campus life.

5. Providing pathways for students to establish their own privacy preferences with digital services, particularly digital courseware. Students should be able to "opt-in" by category of usage (adaptive learning, usage logging, etc.) and the default position should be that digital courseware grants the same degree of anonymity as a print textbook.

6. Ensuring that any contract with a third party that involves the collection of student or faculty data clearly stipulates data use, ownership, and migration terms. Data and information provided, generated, derived, or otherwise created through any service should remain the sole property of the institution or students and faculty themselves, and all uses of this data should require approval. This includes the obvious steps of prohibiting the re-licensing or selling data, but also the use of the data (even if de-identified) in product development, marketing, or profiling.

Engage in Open Procurement Practices

An important area when institutions can assert control of data is through purchasing and procurement processes. These processes should be revisited and revised to ensure that they are transparent, competitive, and fully coordinated across the institution.

The experience of academic librarians entering collections subscriptions with scholarly journal publishers has illustrated the importance of implementing open procurement policies. Commercial vendors often have an advantage in negotiations, since they have detailed information on what each office/department/school/campus is willing to acquire and spend, while the institution itself may not. Sometimes this arises simply from a lack of coordination, but in other cases may be enforced through non-disclosure agreements (NDAs) that prevent institutions from discussing pricing (and also terms and conditions).

Demanding the removal of NDAs from data and data analytics contracts is an important first step that legal offices can take toward open procurement. This will make it possible for institutions to compare pricing terms (and terms and conditions) with their peers, and therefore level the playing field for institutional negotiations.

A critical step is to openly share agreements and contracts with the academic community, making it possible to develop a transparent market price and set of terms and conditions.

Other important best practices in data procurement include retention of data ownership, data migration requirements, perpetual post-cancellation rights to output of data analytics, the preference (or even the mandate) for open source tools and software, the prohibition of disclosure of individual data to third parties (even if anonymized), the notification of subpoenas and government requests for data, and only working with vendors with industry-recognized security certifications.

CDOs or task forces should coordinate with the institution's legal office to establish open procurement policies and develop a framework under which exceptions are permissible. Institutions are likely to be confronted with situations where a vendor may offer a better deal in exchange for an NDA, and it is important to weigh the short term gain versus the long term benefit of a level playing field.

2. STRATEGIC CHOICES

The second category of actions is more complex, since it relates to decisions that will need to be made specifically based on each individual institution's mission, culture and values. It also involves the establishment of an explicit process to determine the position that each institution wants to take in regards to specific issues posed by the collection of data and the deployment of data analytics tools.

SPARC's goal is not to provide answers in this section. Rather, we hope to trigger a broad and thoughtful debate within academic institutions to ensure that these processes are explicitly carried out. The criteria for decision-making in this category are more complex than in the case of risk mitigation actions. It is hard to argue, for example, that the institution should not have a strong privacy policy or conduct a data inventory. The choice between competing actions simply comes down to what is feasible within the institution's resources, culture, and timeline. On the other hand, this section deals with choices that do not have clear right or wrong answers, and where there will need to be a nuanced debate.

It is vital that these debates involve all stakeholders on campus. These debates will be complex and multifaceted, with ethical, legal, economic, and technical dimensions. Many institutions will have scholars and practitioners in these fields right on campus and would be wise to leverage this expertise in structuring the debate. This is very important because there will be diverging views on what the right answers are, but decisions will be more acceptable if reached after using a well-structured approach.

Algorithms vs. Humans

The debate over using artificial intelligence as a substitute for human analysis is already playing out in many parts of society, including the corporate sector. As an example, *Forbes* recently developed a list of 15 business applications of artificial intelligence.[9] Out of these 15 applications, at least five apply to academic

[9] https://www.forbes.com/sites/forbestechcouncil/2018/09/27/15-business-applications-for-artificial-intelligence-and-machine-learning/#2a94d66c579f

institutions as well. Should institutions deploy artificial intelligence tools in the admission process? In recruiting staff? In reviewing and grading non-quantitative exam materials? In identifying potential malicious or unsafe student behavior before it occurs? Should students be able to use accelerated reading software? Should software provide a first line of student support, substituting for teaching assistants? The answer to each of these questions is complex and will vary across and within types of institutions.

One of the early use cases for algorithms on campus was plagiarism checking software such as Turnitin, which has recently sparked debates over accuracy, accountability, and bias.[10] Books such as Weapons of Math Destruction and Algorithms of Oppression have highlighted how algorithms can perpetuate inequities through built-in biases and negative feedback loops. As such, there are significant ethical and legal implications of using algorithms to drive decision-making.

However, there also may be implications of not using them. Algorithms can process information more rapidly than humans and provide tailored services to students (such as adaptive learning) that would be cost prohibitive to deliver through faculty or staff. Also, while algorithms will inevitably contain biases that are built in from the start, machines can analyze datasets more consistently and efficiently than individual humans ever could.

Either way, it is only a matter of time before artificial intelligence further pervades campus decision-making in ways that impact equity, privacy, and allocation of resources. Academic senates, institutional governance boards, and other decision-making bodies should begin a dialogue over the pros and cons as soon as possible. Engaging in this debate in advance will help prepare institutions to be deliberate and strategic about deploying artificial intelligence in ways that are consistent with the institution's culture, values, and risk tolerance.

[10] https://www.insidehighered.com/news/2017/06/19/anti-turnitin-manifesto-calls-resistance-some-technology-digital-age

Quantitative vs. Qualitative Metrics

A second critical set of debates that are necessary is around the metrics that academic institutions use for evaluation. This debate often focuses on the issue of faculty evaluation. Most universities argue that they promote their faculty at all levels through a thorough process of evaluation of each individual, looking both at their intellectual achievements and at their personal contributions to teaching and the life of the institution. Nonetheless, a recurrent complaint is the overbearing impact of publications records and journal impact factors (at least in the disciplines and institutions in which this metric is relevant).

Some institutions are already critically reevaluating how they use quantitative metrics. The University of Ghent in Belgium announced in December 2018 that it would change how it evaluates its faculty. In the announcement, Rector Rik Van de Walle wrote:

"No more procedures and processes with always the same templates, metrics and criteria which lump everyone together" and "The model must provide a response to the complaint of many young professors that quantitative parameters are predominant in the evaluation process. The well-known and overwhelming 'publication pressure' is the most prominent exponent of this. Ghent University is deliberately choosing to step out of the rat race between individuals, departments and universities. We no longer wish to participate in the ranking of people".

More broadly, the the San Francisco Declaration on Research Assessment (DORA)[11] and the Leiden manifesto[12] provide additional valuable frameworks on how to think about research assessment and should be viewed as a valuable starting point on how to transform and enrich the assessment process of faculty and researchers. It may be necessary to also define appropriate metrics to address the complexities of interdisciplinary work, which often are not recognized through traditional metrics.

The debate over metrics also extends to evaluation in other areas of campus life, including academic programs, grading, and return on investment for campus programs. While academic institutions may not be ready to altogether abandon

[11] https://sfdora.org/

[12] http://www.leidenmanifesto.org/

the usage of quantitative metrics to evaluate their faculty, they should consider engaging in a genuine debate on the relative weight that they place on quantitative vs. qualitative assessment, and whether the quantitative metrics they use are representative of the objectives of the institution or just happen to be convenient because they are easily available and easily comparable.

IP Exploitation vs. Knowledge Sharing

Many academic institutions house valuable intellectual property (IP) that is generated through the research activity of its community. While U.S. institutions have been allowed to pursue ownership of inventions based on federally-funded research since passage of the Bayh-Dole Act in 1980, few research universities have successfully reaped rewards, despite the enormous potential value.[13]

The emergence of "big data" and text and data mining has opened up new possibilities for research universities to exploit their IP in profitable ways. Articles and datasets can be mined for insights that can be used by industry, for example to improve the odds of profitable investments in R&D or venture capital. Such activities could generate substantial value for academic research institutions, particularly at a time when the future of government funding is clouded by budget constraints and international competition among academic institutions is rising, driving the need for larger budgets.

On the other hand, vigorous IP exploitation would likely raise a number of ethical issues around partnering with specific industries and companies, as well as concerns that prioritizing IP exploitation could shift resources away from disciplines with less commercial value. Moreover, any decision to exploit data and knowledge for commercial and financial purposes must be weighed against the benefits of Open Data for accelerating the pace of discovery and increasing the integrity of the scientific and scholarly record.

[13] In 2012, the then President of the Association of University Technology Managers testified to congress that as much as 30% of the market capitalization of NASDAQ was driven by academic research. https://www.govinfo.gov/content/pkg/CHRG-112hhrg74722/html/CHRG-112hhrg74722.html

This debate is not necessarily mutually exclusive. For example, it may be possible to maintain an Open Data policy for baseline-quality datasets, while setting up a second flow for datasets that have been processed, cleaned, and standardized for IP exploitation. This structure could provide a “best of both worlds” scenario, where grant funding could support the first flow and commercial services could pay for access to the second flow.

While SPARC is known for advocating for Open Data when possible, we recognize that different institutions can legitimately adjudicate this issue differently. Our goal in this document is not to prescribe answers, but to encourage institutions to hold a broad and thoughtful debate to decide this issue for themselves.

3. COMMUNITY ACTIONS

While individual actions are important, collective actions can have an impact at a greater scale. A third category of actions for the community to consider focus on leveraging a strength in numbers approach, and targeting "big picture" actions institutions to regain and maintain control of their data infrastructure. This category includes a broad range of possible structural solutions to foster an open, competitive landscape for data and data analytics that is aligned with the interests of academic institutions and the communities they serve. Open competition and transparency in this crucial space is essential if the academic community wants to ensure better terms and conditions from commercial vendors, more innovation from different sources, and more truly international/global solutions. By working at scale, these actions have the potential to have significant and long-term impacts, and they are intended to be pursued alongside the important campus-level actions that we recommend in the two previous sections.

Collectively Implement Strategic Practices

The most immediate step that the academic community can take is to strategically leverage its collective market power to change the behavior of commercial vendors. By working collectively in this area, it is possible to create a market where companies must not only compete on price and quality of services they provide, but also on how well they align with community values. Collective actions may happen through existing consortia or networks, state-level coordinating bodies, or new structures.

Common Contract Terms and Conditions

An important first step to consider is to have a critical mass of institutions to demand contract terms and conditions that support a more open and competitive market. In the Risk Mitigation section, we laid out multiple recommendations for institutions to adopt strong privacy policies, data policies, and engage in open procurement practices. All of these steps are important for institutions to take on their own, but the results will be even more powerful if institutions engage in these efforts together.

Broad adoption of common terms and conditions will have a market effect that favors products and services that are in the best interests of the academic community. This includes advantaging Open Source software over "black-box" algorithms and leveling the playing field for community-owned tools to compete with commercial options whenever available.

Buying Time

There are areas of infrastructure where the community can reassert or maintain control before it is lost. This is particularly true in some areas of teaching and student life, where digital tools and analytics have not yet been comprehensively deployed. However, it will take time for viable community-controlled infrastructure to come to market, and therefore an interim strategy may be to buy time. Strategies for buying time include avoiding new services with significant potential for vendor lock-in, putting a hold on new data or data analytics products until some of the actions we outline under Risk Mitigation and Strategic Choices are complete, and reconsidering steps that could accelerate vendor capture of new grey data – particularly "smart" devices and "inclusive access" digital textbook subscription programs.

Build or Acquire Academic Community-Controlled Infrastructure

The most direct path to ensure community control over data infrastructure is to build or acquire it. Currently, the vast amounts of data generated by academic institutions are largely under the control of commercial vendors. While the academic community can and should pursue strategies to ensure these vendors are more accountable, infrastructure that is truly owned and governed by the academic community is best positioned to align with its values and needs. As Bilder, Lin and Neylon pointed out in 2015,[14] "everything we have gained by opening content and data will be under threat if we allow the enclosure of scholarly infrastructures..."

How can the community build or acquire infrastructure?

There are three main approaches, which may be pursued in combination or separately to build or acquire infrastructure. Each of these courses of action requires weighing their trade-offs, and each also requires academic institutions and funding bodies stepping up to invest the resources necessary to assemble viable alternatives to commercial solutions.

[14] http://dx.doi.org/10.6084/m9.figshare.1314859

- **Build from Scratch.** Building assets from scratch would ensure that their design aligns with the requirements of academic institutions and that their governance reflects the mission, values, and goals of the community. It would also likely be a cost effective approach. On the other hand, this approach requires academic institutions to build up very specific and potentially new capacities, and it may take longer to launch products that can compete in the marketplace with commercial vendors.

- **Funding Start-Ups.** Funding start-ups or existing community-owned initiatives would have some distinct advantages over building from scratch, since they come with their own management teams and attendant competencies. In the case of start-ups, a mix of academic and private ownership is possible, allowing the community to seek additional funding across both academic institutions and venture capital funds. On the other hand, venture capital management can be risky and complex. Most important, the goals of academic institutions (launching and running a workable infrastructure) and those of management and funders (maximize financial returns) may come into conflict with each other over the long term.

- **Acquiring Existing Assets.** Acquiring existing assets may be the most expensive option up-front, since sellers may want a premium that reflects their initial risk. On the other hand, existing assets could well provide the fastest way to come to market with the appropriate products and services, as well as bringing in an established user base. However, there may be some issues around interoperability.

The choice between these options is largely a function of the level of funding provided and the availability of potential assets to acquire. As an example, the cost to outright acquire the full set of infrastructure for the scholarly communications process would likely be in the low hundred millions of dollars. To build this infrastructure from the ground up would likely cost less up-front, but would take substantially longer. However, it is worth noting that there are Open Software-based solutions across the entire workflow of scholarly communications that could help the "build" process go faster. The recent "Mind the Gap" work by John Maxwell and team contains a detailed map[15] that provides an excellent blueprint of this particular space for the community

[15] https://mindthegap.pubpub.org/

to consider.
It is important to point out that the financial commitment necessary to build or acquire assets is not a one-time investment.

Any movement towards true community control of infrastructure will require institutions to be willing to invest in digital infrastructure with the same commitment as they currently invest in physical infrastructure.

Like buildings and roads, this will require regular additional investments to ensure that their underlying technology remains up-to-date, stable and sustainable. Emerging initiatives like the Invest in Open Infrastructure (IOI)[16] and the Sustainability Coalition for Open Science Services (SCOSS)[17] could play important roles in ensuring that there professionally managed investment options for the community to consider.

How would the community pay for it?
The first step before any investment would be to conduct a detailed analysis of business plans and potential acquisition valuations. This would allow the community to conclude which options are most attractive in financial terms. Performing this assessment is timely, as many institutions are at a crossroads in reconsidering their current financial relationships with commercial content providers, and consequently, are strategically rethinking their scholarly communications, course material, and infrastructure spending.

More and more frequently, these explorations are taking place within the context of the institutions larger strategic infrastructure investment discussions. For example, one major research institution's President recently tasked the library to analyze how it could cut its spending on scholarly communications by 50%, with the goal of reinvesting its spending into initiatives aimed at radically changing how scholarly communications and infrastructure are managed. If such an approach became a socialized and coordinated effort among leading research institutions in North America or - ideally - globally, it could yield resources approaching the level needed to fund a collective approach to community-owned infrastructure. Likewise, a

[16] https://investinopen.org/
[17] http://scoss.org/

number of institutions have announced six-figure funding for local open educational resources programs, viewing it as an investment that will pay dividends in reducing the overall cost of education for students. As more institutions begin to approach open educational resources as infrastructure for teaching and learning, there is an opportunity to coordinate shared investments toward meeting common needs.

Redirecting current institutional content spending towards infrastructure is an important potential source of funding, but is not the only one. Multiple sources will be required in order to ensure scale and sustainability. Establishing new partnerships between private and public funding bodies and higher education institutions to support community-controlled infrastructure is also critical. There are already examples of funding bodies independently taking an active role in supporting new dissemination platforms (e.g., eLife[18] and Wellcome Open Research[19]). However, it remains to be seen whether joint initiatives between academic institutions and funders can be established and gain traction.

Establish Inclusive Governance Structures

Another approach is to identify or construct governance structures that would allow new kinds of relationships between academic institutions and vendors – commercial or otherwise. While members of the academic community often participate in the "governance" of commercial vendors in an advisory capacity, they rarely have the opportunity to do so from a position where they can exert real operational influence – such as a position on a fiduciary board. If a critical mass of academic institutions were to demand such a role, interesting new opportunities for community-aligned governance could be explored.

It is vital for the governing bodies of infrastructure services to include representation from the communities they serve in order to ensure that management stays accountable to the community's evolving needs. Iterations of this approach have long been a part of the governance of Open Software initiatives, some of which might serve as useful foundations for governance models in other types of infrastructure. Governance bodies should be deliberate about considering which voices are

[18] https://elifesciences.org/

[19] https://wellcomeopenresearch.org/

important to include, and strive for diverse representation across a wide cross-section of factors, including institution type, geographic location, career stage, gender identity, racial or ethnic identity, disciplinary background, as well as other relevant factors.

Leverage Policy to Support Community Control

Updating Federal and State Policy

Another avenue to expand community's control over data infrastructure is to advocate for favorable federal and state policies. Multiple studies have found that the current legal framework is insufficient to prevent the commercial exploitation of student data.[20] In the U.S., the most immediate opportunity centers on data privacy protections, and the primary federal law that governs student data privacy in higher education, the Family Educational Rights and Privacy Act (FERPA). This law was established before the internet, and is limited in both scope of coverage and protection it offers students. Updating FERPA is a clear opportunity to advocate for stronger provisions that guarantee students full control over their education data, protect it from exploitation, and set stronger security standards for vendors.

Another promising option derives from federal-level discussions around potential consumer data privacy legislation, akin to Canada's Personal Information Protection and Electronic Documents Act (PIPEDA) and Europe's General Data Protection Regulation (GDPR). Ensuring that any proposed legislation also applies in the education environment (and carries any additional protections the community deems necessary), would also provide leverage to institutions for maintaining control over critical data.

In both the U.S. and Canada, higher education is closely tied to the state or provincial governments, so legislation at this level could also offer advantages. For example, state law typically governs procurement at public institutions, and open procurement practices could be advanced through a ban on NDAs or other unwanted provisions in relevant contracts. State and provincial governments might also consider introducing regulations that guarantee that students are informed about the terms of use for

[20] https://www.fordham.edu/info/23830/research/10517/transparency_and_the_marketplace_for_student_data/1

digital course materials before registering for a course, or regulating "inclusive access" programs to prevent course materials from becoming an unaccountable student fee.

Antitrust Actions

There are also possible actions that can be considered under existing antitrust law. Both the scholarly communications and courseware publishing markets have become increasingly concentrated in recent years, despite concerns raised within the academic community. The rise of data analytics adds another layer of concern: if the economic model of digital publishing favors oligopolies, that of data analytics favors the rise of monopolies. As our experience of social networks and search engines demonstrates, the future of research and education could end up being defined by singular firms with excessive market power.

The recent decision of the Department of Justice to initiate an investigation into the practices of leading tech companies suggest that the view of regulators over these issues may evolve, particularly in regard to the collection and usage of data. While it is early days, the academic community should monitor how regulators (and the courts) decide on concentration of data and data analytics services and their ties to the provision of other services and – if the situation demands it – initiate or support antitrust actions at the appropriate point in time. The community can also work collectively to take proactive antitrust action by filing comments with antitrust enforcement agencies or working actively to oppose mergers.

Realign Stakeholder Relationships

These community-based actions portend several possible realignments within the academic community and its stakeholder groups that should also be considered as efforts move forward.

Academic and Research Libraries

Within academic institutions, there is need for realignment between libraries and the rest of the institution. Library professionals live and breathe data and information flows every day, and have a unique opportunity to contribute their expertise. In addition, there is a clear need for the senior administration to identify the leaders who will organize these actions, and librarians could well lead some of them. In order to do

so, however, libraries will have to upgrade their project management competencies, and become comfortable mobilizing resources from outside their traditional core activities (for example, from legal, ethics, economics and business experts).

Funding Bodies

The second realignment is within the broader research community. Historically, funding bodies and academic researchers have worked at arm's length, and some of that separation will have to continue (particularly in the grant approval and review processes). However, at a broader level, the issues posed by data analytics portend a much closer and aligned relationship between funding bodies and academic institutions, as they both share some of the objective of keeping research data infrastructure open to competition.

Scholarly Societies

As new partnerships and financial arrangements that ensure greater alignment with community values are considered, the role of scholarly societies might also be reexamined. Historically, the relationship between academic institutions and scholarly societies has been complex and sometimes disconnected, as many academics consider their society to be their primary affiliation before their institution. Additionally, scholarly societies have been perceived, for right or wrong, to be among the least enthusiastic supporters of open scholarly practices, given their concerns over the potential revenue loss in a transition to open access.

In the development of community-owned infrastructure, the relationship between academic institutions and societies might be productively reexamined and realigned to support mutual interests. The development and management of community owned infrastructure requires many functions, including some that are largely dependent on disciplinary expertise that societies alone possess. New kinds of direct fee-for-service arrangements may offer an alternative source of revenue to societies, while supporting direct community control of the communication of research outputs. Similarly, as more institutions invest in open educational resources, scholarly societies are poised to play a potential role as a service provider for vetting or publishing educational materials.

"For example, a number of learned societies have established the Society Publishers Coalition.[21] This is, effectively, a "coalition of the willing" with the aim to establish closer working relationships with academic institutions and funding bodies and may help taking the initial steps to establish more collaborative relationships between the academic community and societies."

[21] https://www.socpc.org/

CONCLUSION

The need for academic institutions to act to retain control of infrastructure, data and data analytics is here to stay. It is critical for academic leaders to acknowledge that data and its uses play a central role in the operations and the future of their institutions, and take control of how it is managed as a strategic asset.

The time to act is now. Many of the actions outlined in the Risk Mitigation section of this roadmap can be taken relatively quickly, and many institutions already have a head start on these processes in response to GDPR or other requirements. Progress can be accelerated by developing and sharing resources that can help meet common needs. There is an important role for higher education professional associations, consortia, compacts, and other community organizations that can provide platforms and channels for disseminating best practices, templates and guides. Similarly, discussions on key issues outlined in the "Strategic Choices" section are already underway on many campuses, and identifying forums to amplify and share these conversations would be both valuable and productive.

Finally, opportunities for galvanizing Community Action abound. SPARC is committed to participating in this process to the extent appropriate, and we encourage community leaders to fully engage as well. Only by working together can we successfully create research and education data infrastructure environment that is open and transparent, that allows and encourages competition, and that operates in a way that is fully aligned with our community values.

November 2019