Executive Summary

More than twenty years into online distribution, the academic publishing industry is undergoing a massive adaptation process. Structural issues are surfacing, leading some publishers to pursue new strategies, with significant negative implications for the academic community.

Elsevier and Pearson (among others) are becoming data analytics companies. Neither company wants to abandon its traditional business: those businesses are very large and failure to sustain them would have severe consequences for the companies' share prices. In addition, content is key to building their data offerings. Therefore, the journal business of Elsevier and Pearson's Higher Education courseware are likely to stay in their respective portfolios. But both companies expect to drive growth through the sale of data analytics services to their customers, rather than through the traditional core business.

By leading the shift of the publishing industry into supplying data analytics, however, these publishers pose challenges for the academic community, which could lose control over vast areas of its core activities (in addition to incurring in significant legal and reputational risks).

RESEARCH

The not-so-gentle giant

Much of the anguish in the academic community towards scholarly communications companies is focused on Elsevier. Elsevier attracts a disproportionate amount of attention within the academic community because of its tone-deaf business practices and its profitability (Elsevier reports a 37% operating profit margin vs. 23% for Springer Nature.) Elsevier would argue that it enjoys record profitability because it

manages itself well and because researchers want to publish in its journals.

The adversarial relationship between Elsevier and academic librarians (and some researchers) is a throwback to management practices that have long disappeared in virtually every other industry. The Open Access (OA) movement, the periodic boycotts of Elsevier by groups of academics, the announcements from European consortia willing to lose access to Elsevier journals rather than agree to demands—all point to a flaw in how Elsevier, and some of its competitors, operate. It is remarkable for a company to face such antipathy from its customers, which also supply the raw intellectual material for its products.

Data analytics will be central to the future of Elsevier for four key reasons:

> **Protect the core journal business.** Elsevier must recognize that Open Access is a significant threat to its economics. We estimate that Elsevier revenues per article are about £3,100, well above the estimated industry average Article Processing Charge (ACP) of £1700 to 1800. This means that Elsevier faces a potential significant decline in revenues if a full transition to OA takes place. Elsevier may not cut costs enough to compensate for lower revenues, as - with estimated costs per article of about £2,000 - a 10-15% cut would only save about £200 to 300. In this scenario, the operating profits of Elsevier would be wiped out almost completely, as both revenues and costs could converge around the £1700 to 1800 mark. The best way for Elsevier to protect its profitability would be to gain share among the highest impact journals which can and will support higher APCs, effectively nearing, as much as possible, its current average revenues per article. This, in turn, would be easier if Elsevier could win larger numbers of high impact submissions from authors. In this scenario, the more "high impact" articles Elsevier can publish thanks to early insights into research likely to have high impact (thanks to data analytics), the better.

Improve how Elsevier runs the journal business. Elsevier holds one of the largest repositories of data on citations and readership through its various databases – from Scopus and Science Direct to Mendeley, SSRN and Bepress. These databases enable Elsevier to gain insights which will affect its competitive position in the future, often with an advantage which may be measured in years. Unsurprisingly, Elsevier is the only leading publisher which does not participate in I40C, the Open Citations project backed by virtually the entire industry). Elsevier can identify which researchers are likely to grow into future leaders in their respective fields, which areas of research are likely to evolve into the next growth themes, and where interdisciplinary studies are likely to develop. Elsevier can use these insights to cultivate rising luminaries by adding them to editorial boards early in their careers or to launch new journals ahead of competitors. In addition to these strategic opportunities, there are more tactical ones that may allow the company to run its operations better and at lower costs. Smaller publishers have no possibility to replicate these advantages, because Elsevier has a sizeable lead in assembling data and because there are few assets left which can be acquired to match Elsevier's capabilities.

Sell insights to universities, funding bodies and governments. The insights which we discussed in the previous paragraph are also valuable to universities. Elsevier and Clarivate (the owner of Web of Science and a number of research productivity and intelligence tools) as well as smaller competitors, such as Academic Analytics, already sell tools aimed at delivering research intelligence along several dimensions. Some of the tools are still primitive (several interviews suggest that Clarivate is ahead of Elsevier in terms of quality, but Elsevier is working to fill the gap) but they will improve over time. In addition to driving resource allocation, these tools can also affect other core processes of universities, like hiring and tenure. Once deployed, customers may find difficult

to discontinue usage of these tools: the experience of the corporate sector is that once users rely on specific data tools, they find it difficult to abandon them. This means that these services benefit from strong pricing power, and Elsevier is uniquely positioned to offer them because it also has access to content and underlying data sets.

Sell insights to industry or the investment community. This opportunity is the most speculative, but also the most valuable. Elsevier controls a massive amount of potential intellectual insights, and, increasingly, data. It would make sense for Elsevier to capitalize on its access to academic research by partnering with industry to exploit insights, or by establishing joint ventures with venture capital firms to improve returns. Elsevier may only think today of monetizing this value by licensing software to mine articles and data repositories. On the other hand, Elsevier may realize that it could partner with others to exploit these insights.

Clarivate – the natural competitor

A focus on improving their economics through industry consolidation has absorbed the management of Springer Nature, while Wiley has chosen to invest resources in improving its digital operations through the acquisition of Atypon. As a result, Clarivate appears today as the only company in a position to challenge the data analytics business of Elsevier.

When Thomson Reuters (TRI) owned the Science business, it kept it separate for many years from its other assets in the space (healthcare and IP). Effectively, TRI was pursuing a strategy narrowly aimed at serving libraries and academic institutions with tools which would be "neutral" in the eyes of its customers and of the scholarly communications industry. After the disposal of most of the healthcare assets in 2012, TRI consolidated Science and Intellectual Property into one group, in order to offer products which bridged research and commercial applications.

It is unclear whether the new owners support this strategy. The company has made two acquisitions aimed at modernizing its research offering, but neither upends the strategy as a "neutral" data analytics supplier to researchers and librarians. Still, integration between data analytics and intellectual property remains a significant

integration between data analytics and intellectual property remains a significant opportunity, as it can lead to serving universities seeking deeper ties (and funding) with industry and industry looking to support relevant research.

The issues

The three leading research data analytics vendors (Clarivate, Digital Science, Elsevier) market a wide array of research intelligence tools, raising substantial questions.

All three companies have targeted customers inside and outside the academic community. This raises the first question for universities: is it in their interest to share additional data with companies that will feed the data to funding bodies and governmental decision-makers?

Second, using data is obviously important when assessing productivity and impact, but no algorithm can capture the complexity of the judgment that universities are asked to exercise. Well-funded institutions can afford the resources to evaluate every researcher. On the other hand, institutions with limited budgets or those wishing to demonstrate objectivity may be tempted to give disproportionate weight to analytics tools. Also, each institution is different in how it weighs quality, performance, and impact. Using standardized, non-transparent algorithms could lead to decisions in conflict with the values of any specific institution.

Third, what is the use of data within these companies once it is made available by research institutions? In particular, what degree of privacy is afforded to university and researcher data once it is uploaded in order to utilize the productivity tools made available? Is this data fed into other businesses within the same company? Is it made available or resold to third parties? How will a commercial vendor respond to a government request for disclosure?

Fourth, some of these businesses effectively text and data mine materials submitted by researchers. What is the licensing agreement underlying these relationships? Do other researchers enjoy similar access rights? One of the issues brewing in the controversy around the boundaries of IP is the limitation demanded by many publishers to the use of third party software for the purpose of extracting insights.

This is not a comprehensive list of all the issues posed by the provision of data tools and services outside of university libraries. We expect the list of issues to grow over time, as more people weigh in from their vantage points on the basis of their experience and concerns.

TEACHING AND STUDENT ACTIVITIES

Higher Education textbook publishing is affected negatively by trends which are structural, rather than cyclical, in nature. Numbers were startling: in 2016, the major publishers posted revenues declines ranging from -10% to -15%, followed by another -3% to -6% organic revenue decline in 2017. This continued decline is leading publishers to support the switch to fully digital courseware products, under the umbrella of "inclusive access".

In its simplest form, inclusive access is a flat-rate subscription to a publisher's catalog of digital materials. Students can lower their courseware costs by replacing individual purchases with an annual subscription. Consortia like Unizin, which negotiate across all publishers, act as aggregators to negotiate lower rates. From the publishers' perspective, the combination of lowering costs and adding users means that their economics are equal or better.

In an overly simplified view, inclusive access is a win-win situation for both publishers and students, leading to equal or higher profits for the companies and lower spending for students. However, things are not as easy as they appear, particularly for students. All the students who lower their spending by reselling books, renting them, sharing them or by checking them out of libraries may not be better off (and many will be worse off). Most important for the purposes of the issues raised in this document, once students transition to digital materials they enable both their institutions and the commercial vendors to collect vast amounts of data on them: their physical location when they use them; their study habits; their learning profile; granular knowledge on their performance. This poses significant privacy issues, and—potentially—legal liabilities which may become, at some point, very grave.

For many years, Pearson has been the only large, publicly-traded courseware company. Management has been very vocal, at least with the financial community, about the declining value of print-based courseware, to the point that many investors interpreted the message as "the long term value of the print textbook

business is zero". Whether management meant to convey this message is irrelevant. Effectively, Pearson has educated the financial community that only digital courseware and data analytic services will have value in the future. Pearson is now committed to driving the fastest possible shift to digital courseware.

Pearson has explicitly highlighted, in some of its investor presentations, how data analytics can affect outcomes for students. Management believes that digital tools are not just aimed at increasing faculty productivity and lowering costs. Pearson's strategy indicates its belief that a combination of student profiling through adaptive learning tools (which identify the best way to teach to different segments of students) and reporting will transform schools with very large enrolment numbers, where it is less cost effective to monitor individual students.

So many issues…

The decline of print courseware is forcing publishers to move towards digital content. Once content is digitized, however, it opens up vast possibilities to collect and analyze user data. Universities must consider the ramifications of collecting and monetizing such large amounts of sensitive data.

Digital products include not only courseware, but also assessment, immediate feedback, content customized on the basis of the learning profiles of students, homework tools, standalone platforms or Learning Management Systems, lecture capture, and more. Evidence suggests that these systems capture massive amounts of data about student (and faculty) behavior that go beyond what is necessary for improving student outcomes. Digital tools can determine a student's learning "profile", where and when students access content or complete homework, what resources are accessed to complete tasks, how long it takes to complete exercises, etc. This data—if hacked, resold or surrendered to governments—can be used to classify students, screen them for employment or access to graduate education, infer their political views and map their personal networks. While many students today have a choice whether to acquire print or digital courseware, the trends towards restricting them to digital products are unmistakable. Digital courseware lowers costs for publishers and enables the sale of data analytics, while helping universities increase productivity.

Another area of concern is what is contained in the Terms of Use of digital textbooks, which are typically signed by the student rather than the institution. Yet, the decision of a faculty member, an employee of the institution, to assign a digital textbook makes it a foregone conclusion that students must accept a publisher's terms in order to access their course materials. It is standard for digital products' terms of use to allow the provider to change the terms at any time without notice, possibly retroactively. Faced with increasing financial pressures and tempting opportunities to monetize data, it seems unlikely that publishers will be able to resist.

The risks to student and faculty privacy are significant, ranging from hacking to unmonitored resale of data to third parties. Could publishers find themselves selling student data, even inadvertently, to the next Cambridge Analytica? Will publishers resist government requests for data? Would universities resist requests for selective data from prospective employers?

The algorithms themselves used by the publishers are not transparent, which also raises questions. For example, how do adaptive learning tools classify each individual's learning profile? Are all the profiles valued equally, or are systems ranking students on the basis of perceived abilities and learning styles, handicapping some even before they complete a class? Numerous examples in recent news illustrate how algorithms can be influenced by the unconscious biases of those who design them, which can result in unintended discrimination. In the higher education context, where algorithms are increasingly trusted with important decisions, the lack of transparency raises both ethical concerns and potential legal exposure.

Finally, it is worth noting that publishers are not the only vendors of data analytics to universities. Many colleges are using data analytics (to various degrees) in the recruiting process, and there are many fairness questions raised by the use of non-transparent algorithms. Are algorithms perpetuating, even involuntarily, biases based on geography, occupation, likelihood that students or their families turn into

donors, etc.?

PRELIMINARY RECOMMENDATIONS

This document aims at outlining issues posed, rather than proposing detailed solutions. However, it is useful to highlight some of the possible responses available to the academic community. We believe that responses will have to be tailored to individual universities, although some would likely require collaboration among institutions. For any one institution, the key drivers will be a mix of cultural elements (e.g. How centralized or decentralized are decisions?) and financial considerations (e.g. What is the expected impact of data analytics? Are significant revenues from Intellectual Property desirable? Which companies are acceptable partners?). We will offer choices rather than prescribing a single course of action.

We think that there are two different sets of responses. The first set is aimed at risk mitigation. These are actions aimed at protecting institutions from the possible consequences of deploying data analytics tools and collecting larger amounts of data. These actions may include revising existing data policies and strengthening compliance mechanisms, and adopting "open procurement" processes. These policy shifts could mandate that institutions maintain ownership of data, that algorithms are made fully transparent, that results from their usage are portable, that data is not resold to third parties, that sensitive data is not turned over to government agencies without first resorting to the courts, that contracts are made publicly available, etc. Risk mitigation may also require appropriate organizational support. Some institutions may opt for coordination groups, while others may decide that the senior administrators may identify individuals tasked with revising data policies, monitoring execution, and helping individual offices negotiate and/or adjudicate conflicts of interest.

A second set of responses will require individual institutions to decide among trade-offs. For example, using data analytics to drive the admissions process is less expensive than hiring a large staff. Different institutions may choose different levels of human intervention to balance budget and cultural priorities. Similarly, different research institutions may choose different balances between commercial exploitation of intellectual property and funding of research with limited commercial upside. Some universities will support community solutions to these issues; others

upside. Some universities will support community solutions to these issues, others will be skeptical. In a separate document, we will offer an array of solutions and help individual institutions consider these issues.

Alternatively, universities may consider pursuing antitrust cases against the largest suppliers. Whether such a course would succeed is unknown. To the extent that research intelligence tools are based on platforms for research dissemination and that there is a similar integration between digital courseware and data analytics tools among courseware publishers, it may be possible to argue that that there should be remedies. Whether the academic community has appetite for such strategy remains to be seen.

Finally, institutions could respond by adopting community-owned solutions. This could take many forms – acquire existing assets, sharing the costs of maintaining existing institutions-owned tools, or jointly developing new ones. We think that this approach has much merit and should be carefully studied.

We believe the academic community still has time to act. Taking action is urgent because terms and conditions are being established right now, but taking stock of the situation, asking the right questions and choosing the right course of action can still prevent the academic community from finding itself, against its interests, unintentionally obliged to follow a path leading to losing control of its future.

<div align="center">*</div>