

Open data and metadata - opportunities, risks and possible actions

JULY 27, 2020

Presenter:
Claudio Aspesi

Who am I? Why am I here?

- Lead author of several SPARC reports on data analytics and academic institutions over the past 2 years
- Started involvement in academic publishing as the Senior Research Analyst – European Media at Bernstein
- Before that, 20 years of experience at McKinsey and then as Head of Strategy for EMI Music

Today's discussion

Research data is at the core of what universities do. Its value to researchers is, of course, paramount - and open science offers significant benefits to the scientific community. But this data, and the attached metadata, are increasingly valuable for third parties as well. We will discuss

- How research data and metadata increasingly overlaps with all the other data produced by academic institutions
- How it is becoming increasingly valuable outside the academic community, and how it could become even more valuable in the future.
- The challenges posed by the collection, analysis, synthesis and preservation of data and metadata
- The principles that should be used when analysing data and using Artificial Intelligence
- What are the ethical questions that need to be addressed.
- What are possible steps to take to address all these issues.

Open Science has many benefits

- Preservation and retrieval of data sets becomes easier
- Open peer review ultimately improves credibility or research processes
- Higher discoverability and visibility of the work leads to stronger ties with the academic community
- Reputational gains lead to higher probability of publication, higher citation count of articles

Research data is increasingly becoming open to third parties...

- Data sharing policies are spreading – recently, the U.S. Federal Government has mandated open data, Horizon 2020 in Europe is requiring OpenResearch data and others have established some form of open data requirement
- The *State of Open Data Report 2019* reports that 79% of 2019 respondents were supportive overall of a national mandate for making primary research openly available
- Practices and protocols for opening data are being established

...but, at the same time, massive amounts of “gray data” are being generated and collected with little or no transparency

- Activities like teaching, studying, checking reference materials in the library, etc. are all leaving digital trails
- Most of this “gray data” – as Professor Borgman at UCLA calls it - is being collected with no requirement for openness (and, in fact, with some layers of privacy protection).
- This “gray data” is, or could be, someone else’s research data, and protocols need to be developed for their use in research – but also on its uses today outside of research.

Research data is extremely valuable

- In a hearing before the Subcommittee on Technology and Innovation of the House of Representatives in 2012, the then President of the Association of University Technology Managers indicated that NASDAQ estimates that 30% of its value is rooted in university-based research results.
- Based on last Friday's prices, that is about \$4.1 trillion

What emerged during the SPARC project

The commercial vendors' strategy is **to lock up exclusive control** of higher education infrastructure, and with it the data it generates/stores.

- Data and data analytics affect all facets of academic life and will play an even larger role in the future.
- Algorithms (for example, impact factor, university rankings) already affect academic life.
- The use of data analytics affects the core mission of academic institutions and potentially create severe, unintended consequences.

Implications

The dangers of proprietary algorithms are significant. Exclusive control of data allows commercial players to create “black box” algorithms that can:

- Influence selection of the students.
- Classify students based on their learning style.
- Evaluate the “productivity” of faculty.
- Identify “contributors” to research “success”.
- Decide funds and staff allocation among departments, labs, and individual projects.
- Drive external funding of research.

While many of the purposes of data collection and analysis are benign, there are risks

- We are not luddists. Without data, biases and injustice thrive –we do not oppose data analytics and AI in principle
- Algorithms and metrics are designed by humans, and carry risks, from *bona fide* errors to biases. These issues have been reported in books like *Algorithms of Oppression* and *Weapons of Math Destruction*.
- Warnings, however, are ignored. There is some discomfort around consumer uses (search engines, social media), but data analytics/AI are still viewed as “cool” and “leading edge”.
- There is even less attention to the market structure of data analytics and AI. These businesses become monopolies or very concentrated oligopolies because of scale and scope economies.
- If academic data analytics become monopolies, biases will be magnified by lack of diversity.

Three recent trends to watch

The COVID-19 crisis has reinforced the attractiveness of some of the emerging strategies of commercial vendors:

1. Continued mergers/consolidation
2. A rising involvement of commercial vendors in research assessment
3. The emergence of even “Bigger Deals” that bundle research assessment with research dissemination and student/faculty assessment with courseware

However, positive signs also exist

- More libraries are pushing back on transformative deals and closely scrutinizing terms and conditions.
- Opposition to the Cengage/McGraw-Hill Education merger was broad and involved a wide array of parties.
- Concerns about data and privacy in relation to academic institutions continue to gain prominence.
- Institutions (including the UC System) are actively embracing SPARC recommendations.
- More actions from faculty and other stakeholders are likely to come soon.

Thinking About Solutions...

We need to explicitly distinguish **metrics** from **algorithms**.

- Metrics (***what* to measure**) are the exclusive responsibility of academic institutions.
- Algorithms (***how* things are measured**) can come from many sources.

These two categories pose different challenges and need to be addressed separately.

Three Categories of Solutions

- Risk Mitigation: activities that we would expect virtually every institution to perform in order to minimize the risks posed by data analytics.
- Strategic Choices: activities which require a full debate among stakeholders within institutions in order to choose a course of action.
- Community Solutions: activities which require coordination and cooperation across multiple institutions.



A framework for action

Open Data Infrastructure: A Roadmap



Actions Based on Mission, Values and Principles

- Institutions should develop principles to help guide their decisions around research and education infrastructure, in order to ensure that these services are transparent, open, and fully support the mission and values of the institution, as well as the scholarship.
- SPARC and COAR have produced a set of seven principles for scholarly communications services that may serve as a starting point.

<https://www.coar-repositories.org/news-media/good-practice-principles-for-scholarly-communication-services/>

How do the SPARC principles translate in the data analytics context?

Principles of Data Analytics Usage*

- **Transparency.** Open source software, disclosure to enable testing for biases, auditing and evaluation requirements, etc.
- **Strong privacy protection.** Consent, control over the use of data, right to erasure and correction, right to restrict processing, etc.
- **Accountability.** Remedy for automated decisions, ability to appeal, etc.
- **Equity.** Identification/correction of errors/biases, fairness, environmental impact, etc.
- **Human control.** Opt out of automated decisions, human review of recommendations, etc.
- **Customization.** Definition of non-standard reports, development of tools for a subset of users, etc.
- **Governance.** Effective input from all stakeholders, independent review mechanisms, etc.

* Based on *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI* published under the auspices of the Berkman Klein Center at Harvard University

Recommended Actions: Risk Mitigation

Sample Action 1.

Higher coordination on campus, to treat *all* data – research, educational, administrative – as a strategic asset is essential. The establishment of a coordinating mechanism (Chief Data Officers, Data Council, Data Steering committee, etc.) should be institutional priority.

Recommended Actions: Risk Mitigation

Coordinating bodies should consider questions including:

- What problems/opportunities is data expected to address?
- How should different categories of data (research, student, administrative) be treated to reflect their differences?
- What data can be shared with third parties?
- Who should maintain ownership for the data itself?
- What rights should the institution secure with regards 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?

Recommended Actions: Risk Mitigation

Sample Action 2.

Institutions should review and revise data policies to address strategic data management in addition to protect from technical risks.

Why Revisit Institutional Data Policies?

- Existing policies tend to be **technical** and **tactical**; **not strategic**.
- They are largely focused on minimizing the risk of **unauthorized data access** (password protection, network protection, classification of data by level of mandated security, etc.)
- Data policies should also address the **authorized access to data** (why, when, under what legal terms and conditions, with what level of supervision), including resolution mechanisms in case of conflicting goals within academic institutions.

What should Data Policies cover in addition to security?

Goals of data and data analytics usage

- Support of institutional mission (leadership in research and teaching, merit-based advancement, social equity, etc.)
- Improvement in transparency of decisions
- Better understanding and fulfillment of needs of all categories of stakeholders (faculty, students, staff)

Principles of data analytics usage

- As discussed earlier, concrete principles that should guide data activity

What should Data Policies cover in addition to security?

Governance mechanisms

- Supervision
- Coordination and participation of stakeholders
- Institutional accountability

Control responsibilities and mechanisms

- Auditing of activities
- Documentation of compliance
- Periodic policy impact assessment

How to manage research data?

- **Research data, in general, is different from "gray data".** Its use and preservation pose different problems, since its collection and management are typically entrusted to researchers or - in some disciplines - to community repositories.
- **Even just preserving research data and the accessory "knowledge" and infrastructure to use and interpret it is challenging.** We are much more concerned about data that has a long shelf life than about data that is difficult to access, preserve or reuse.
- **Research data management should be left to researchers and their communities,** with librarians providing much needed support to ensure that valuable data that could be reused is preserved and remains reusable.

Recommended Actions: Risk Mitigation

Sample Action 3.

Institutions should review and revise **procurement procedures** and **contracting terms** to ensure appropriate ownership and use of data.

Recommended Actions: Risk Mitigation

Consideration of procurement procedures and contracting terms and conditions should include:

- Early involvement of legal support
- Demand for transparent disclosure of contractual terms and conditions (no NDAs)
- Retention of data ownership
- Perpetual post-cancellation rights to outputs
- Preference for open source solutions when feasible
- Prohibition of disclosure of individual data to third parties
- Notification of subpoenas and other governmental requests

Recommended Action: Strategic Choices

- This category encompasses resolving issues which require a full debate among stakeholders within institutions in order to choose a course of action.
- Solutions/Actions will be **(rightly)** highly individualized to reflect individual institutions mission, values, and priorities.

Recommended Action: Strategic Choices

Issues that require this kind of discussion/debate include:

- Software vs. human balance across many functions; “use data, do not rely on algorithms”
- Qualitative vs. quantitative metrics
- Intellectual Property exploitation vs. knowledge sharing
- How to deal with software predicting future behavior
- Research data preservation vs. privacy

It's imperative that these discussions involve all campus stakeholders and draw on expertise of specific disciplines (law, business, economics, ethics, etc.)

Recommended Action: Community Actions

While the first two categories focus on actions that individual institutions might take, the third category focuses on opportunities for collaboration/collective actions.

- These larger-scale actions can be focused specifically in either the research or educational domain, or be structured to encompass both.
- They may be organized around areas as diverse as financial investment, legal action, policy/regulatory action, or construction of new governance structures.

Recommended Action: Community Actions

For community actions to work, a mindset change is needed.

- Institutions must must be willing to invest in developing and continuously upgrading digital infrastructure on the same scale as physical infrastructure.
- The role of the library in managing data and data analytics is of central importance to achieving the core mission of the institution, and must be recognized.
- Broaden the definition of ‘community.’ The interests of funders, state higher education commissions, learned societies align more closely with those of academic institutions than with corporate interests and should be brought in as partners.

Recommended Action: Community Actions

Example 1: Establishing **community-control** of research communication and assessment **infrastructure** is critical. Collective actions can be organized around:

- Collective investment in **building** open infrastructure.
- Collective investment in **buying** infrastructure.
- Development of ways for community to assert control of infrastructure through new **governance** models.

A combination of these approaches can be deployed ensure the academic community controls their own destiny.

Recommended Action: Community Actions

Example 2: For student and learning data, the prospects for community-owned infrastructure are more complex. However, collective actions can have an impact, including:

- New types of consortia allowing universities to obtain both lower pricing for students and better terms and conditions.
- Interventions to slow widespread adoption of damaging commercial pricing models for learning materials.
- Pre-emptive regulatory challenges to highlight risks posed by commercial consolidation in higher education data market.

A time for radical change

- This is the moment to make substantive progress toward building a more equitable and open future.
- Data analytics services are in demand because academic life has become a race to secure funding.
- The global research community is embracing new forms of scholarly communication in the fight against COVID-19. This culture of collaboration is opening the eyes of skeptics to the value of sharing knowledge rather than locking it.

A time for radical change

- University and college leaders will have to decide whether sharing infrastructure, programs, labs and departments may be better than closing them altogether.
- These changes will not come in one week or in one month. But it is time to think about collaboration and sharing, rather than competition, as a basis for rebuilding the academic community.

Questions?