Proxying the Data Body:
Artificial Intelligence, Federated Identity, and Machinic Subjection

Sam Popowich
University of Alberta
Sam.Popowich@ualberta.ca

Abstract

Academic libraries have recently seen a shift from self-management of user-authentication of licensed resources themselves, to cloud-based implementations of "federated identity" technologies. Such technologies aim to solve the problems of fragile access to licensed resources while also better protecting publishers' intellectual property. However, federated identity systems raise a host of issues regarding privacy, surveillance, machinic subjection, and algorithmic governance. This paper traces the development of federated identity systems out of earlier authentication processes, shows how such systems use artificial intelligence techniques to create a trackable "data body" for each student, and then analyzes this whole procedure through the critical theories of Maurizio Lazzarato and Bernard Stiegler. In conclusion, the article argues that the emergent nature of the "data body" creates ambiguity between the hyper-control of contemporary technologies and the possibility of resisting them.

Introduction

One of the primary roles of library information technology departments is to connect disparate computer systems to enable search, discovery, and access to licensed resources by approved members of the library's community. The individual systems so connected are often of different generations, produced by different companies, or developed in-house (Varnum, 2016). While these systems are more or less interoperable, significant effort goes into making sure they can communicate with each other and that the data they exchange is mutually intelligible. The requirement on the part of the owners or publishers of licensed resources to protect their intellectual property and to ensure data integrity and security creates a tension with the library's desire for openness, ease of sharing and reuse, and efficient user-access to the resources themselves. This tension makes sites of social, cultural, and political struggle of systems and the data they contain, struggle mainly between the values of libraries and of private corporations who own (often through the uncompensated labour of academics) the scholarly articles and other resources used by the education and research community. However, it is also part of a broader socio-political struggle over access to education, private property, and hierarchies of power and marginalization operative within broader society. To engage with this struggle, Bess Sadler and Chris Bourg have called for an "explicit feminist agenda" to overcome "the same processes of exclusion and marginalization that have always influenced libraries – and therefore scholarship" (Sadler & Bourg, 2015, p. 1). Furthermore, given that library systems mediate the products of

http://ejournals.library.ualberta.ca/index.php/JCIE
doi 10.20355/jcie29410

scholarly dissemination and their readers, discovery, search, and access systems are a core component of the scholarly communications lifecycle, and are implicated in the problems associated with scholarly publishing more generally (Moore, 2019). But the problem is wider than that, as Oliphant and Brundin (2019) argue: while the application of data to higher education in the form of learning analytics is aimed at training and optimization, that application loses sight of education itself, which is "underpinned by the rigor of theory and philosophy, but also aims to develop social consciousness and critical thinking and acknowledges the importance of experience and experiential learning" (Oliphant & Brundin, 2019, p. 19). This article attempts to situate the particular phenomenon of federated identity systems within this larger contested terrain.

Trying to solve the problems of interoperability between systems, data integrity, and resource security has led to an ecosystem in which the entire network is held together by a series of intermediate computer systems (proxies, link resolvers, identity and authorization databases), each of which is a point of fragility where the operation of the network as a whole can potentially break down (Popowich, 2017). Such breakdowns are reasonably common, and are immensely frustrating to librarians, students, and educators. There is, then, a socio-technical impetus to innovate within this space and to develop new and robust software solutions which can offer seamless access to licensed resources, an impetus which is part of a higher-level logic of technological innovation within libraries (Popowich, 2018). Within scholarly communications, technical innovation is also currently the site of struggle: sites like SciHub, for example, which seem, from a user's perspective, to effectively solve the discovery, access, and interoperability problem, drive publishers and vendors to implement new solutions to satisfy the needs of users while also protecting their intellectual property (Anderson, 2019).

Since 2011, the socio-political conjuncture in which library systems are embedded has shifted to an "Age of Big Data" (Fuchs & Chandler, 2019) in which social media companies, analytics/metrics providers, and other agents of "platform capitalism" (Srnicek, 2016) seek to maximize the amount of data that can be associated with and harvested from platform users, in order to predict behaviour and desires, proactively satisfy consumer wants, and effectively surveil and discipline civil society. These processes are being adopted within the educational sphere perhaps even faster than elsewhere, for example in the use of facial recognition software in schools (Simonite, 2019). Shoshana Zuboff has characterized the current conjuncture as one of "surveillance capitalism" (Zuboff, 2019), but where Zuboff, like Srnicek, sees a radically new, *sui generis* form of capitalism, it is more likely that surveillance and platform capitalism simply continue and extend tendencies already present within capitalist technological imperatives (Dyer-Witheford, 1999; 2015). Looked at from that perspective, federated identity is not only an opportunity to overcome the limitations of library discovery systems, but is also an extension of the technological "society of control" (Deleuze, 1992) into the educational sphere.

What is new, however, in the age of Big Data and surveillance capitalism, is the desire on the part of various actors to capture and harness all the data potentially available from library systems: data on students, faculty, and researchers, their reading and study habits, their location,

doi 10.20355/jcie29410

and their academic success, as well as correlations with other available data (Jutting, 2016; Sibbald & Handford, 2017). This raises important ethical issues for librarians (Jones & Salo, 2018; Jones, 2019a; Jones et al., 2020) and places libraries and the academy squarely in the context not only of surveillance capitalism but of the quantification of subjectivity itself (Moore, 2018). The need for new technological solutions to the fragility of access coupled with the desire to harness user-data – especially student data – has led to the rise of a new set of software protocols which can do both. In this sense, then, the academy – and the library – must be aware of and understand the ways in which it contributes to the cultural, social, and ideological reproduction of a set of values and subjective commitments (Popowich, 2020). This awareness entails particular ethical commitments to either knowingly reproduce those values, or challenge that reproduction in the name of emancipation from the overdetermination of high-technology capitalism. As Nicholson, Pagowsky, and Seale (2019) put it recently, "the focus on what is quantifiable and measurable in the present moment in order to construct a known future erases structural inequities, individual histories, and difference" (p. 66). Libraries are caught between a legitimate fear for their students' privacy and a desire to improve the student experience of discovery and access systems, but they are also embedded in dynamics of socio-political reproduction (Bales, 2015, p. 112-113; Popowich, 2019, p. 151-156). This double-bind is the concern of the current paper.

Access and Authorization

There are various ways software systems (databases, publisher sites, etc.) that can determine who should have access to licensed resources. For example, each vendor could require the creation of an individual user account for their site, as is common with e-commerce sites like Amazon. While this is straightforward enough from the vendor side, the number of vendors active within the publishing space makes the burden on the student too high; each student would potentially be required to create dozens of user-accounts in order to have access to all the licensed content they need over the course of their academic career. As a result, most libraries tend to use an authentication mode based on "internet protocol" (IP) addresses, in which the library reports to each vendor a block of computer addresses associated with the university and the vendor allows access from any one of those IPs. This works well for direct access from on-campus, but it requires another solution for off-campus access (Kilzer, Black & Muir, 2008). Two solutions which have been widely used are 1) for off-campus users to connect to the campus network (thus using a campus IP) using a Virtual Private Network (VPN) system (Varnum 2019, p. 63) and 2) routing the traffic from the user's off-campus IP address through an through an on-campus IP address by using a proxy server (Webster, 2002). The main difference between the two is that VPN access covers all network access by a user's machine, while proxy access only covers traffic that takes place through a web-browser. Outside the library world, a VPN access has become more common in order, for example, to protect users' privacy on the web (Newman, 2017), but VPN usage has declined within universities, due to the responsibility of maintaining a VPN (on the campus side) and of installing and maintaining VPN client software (on the user side). Conversely, while in the past proxying browser traffic was once a fairly esoteric networking process, the preserve of system administrators, the rise of platform capitalism (social

doi 10.20355/jcie29410

media and streaming services, for example) brought with it a need for users to mask their IP to protect their privacy, to circumvent intellectual property laws, or to bypass censorship or geographical restrictions (Montgomery, 2015; Earle, 2016).

The most widely-used proxy in the library world is EZProxy, developed in 1999 by Chris Zagar and purchased by OCLC in 2008 (OCLC, 2008). While it is a core piece of library software, it remains more or less unchanged since its inception. Legacy code is notoriously hard to maintain and update, so EZProxy has not benefited from the addition of post-Web 2.0 features like analytics and tracking capabilities that have become core concerns of platforms and vendors in the library space. While the proxy server necessarily authenticates through a user-database like a Lightweight Directory Access Protocol (LDAP) server, LDAP *authorizes* user access but does not automatically *identify* a user, and while the proxy server logs store a user's ID along with other access information, this data is managed according to the library's security and privacy policies, and is not transmitted to a vendor. In cases of alleged abuse or infringement, such as high-volume downloading, librarians can track down and associate an IP address with a login ID, but this information is never handed over to a third-party.

In order to deal with the fragility and difficulty maintaining a proxy server, as well as to gain the analytic capacity of modern systems, various solutions have been proposed over recent years to move libraries away from IP-based authentication without requiring students to create individual accounts on each site. An early tool for attempting to manage this problem in a new way was the implementation of the Security Assertion Markup Language (SAML) in software like Shibboleth and OpenAthens. These work as single sign-on systems, authenticating once against a user-directory like LDAP and then passing authorization credentials on to any subsequent system that requires them. More recently, a prominent alternative way of approaching the problem has been proposed known as "federated identity" systems (Tay, 2017). The two frontrunners in federated identity are Google's Campus Activated Subscriber Access (CASA) system, and RA21, which is developed jointly by STM (a trade association of academic publishers) and the National Information Standards Organization (NISO). Both CASA and RA21 seek to make access to licensed resources more seamless and effective by managing authentication and authorization in the cloud. While this is a laudable goal in and of itself, it is presumed that such solutions will also discourage piracy and bring users back from sites such as SciHub (Hinchliffe, 2018; Schonfeld, 2018). However, both CASA and RA21 have the capacity to build what Phoebe Moore (2018) calls an "identity proxy or data double," and what, following Antoinette Rouvroy (2013), we might call a "data body" (p. 157; Cheney-Lippold, 2017, p. 115) for each student, raising fears among librarians for student privacy, as the data body can potentially be used for tracking and surveillance.

The Data Body

The data body is not simply a digital representation of the individual student, composed of the myriad traces left in digital systems inside and outside the university's networks; it is also the networked multitude of users which together form the "academic body." Philosopher of technology Bernard Stiegler (2016), whose work we will return to below, argues that the "automatic society of hyper-control is a society founded on the industrial, systemic and systematic exploitation" of such digital traces, and he concludes that "all aspects of behaviour… come to generate traces, and all traces become objects of calculation" (p. 28). In other words, academic behaviour, activity, and decisions, as well as all the passive attributes to which people are subject, such as grades (for students) or student-evaluations (for faculty), form the bedrock of the quantified exploitation by capital of the data body itself (see, for example, Hartman-Caverly, 2019).

The traces left by "residents" of digital space is a core element in the desire to track data bodies, often for justifiable reasons. For example, the creation of a data body for each student can, it is argued, allow universities to connect library use with student success (Tenopir, 2013, p. 273) by being able to link how much and in what ways a given student uses library resources with the grades they end up getting. Additionally, in a recent book on data-driven decision-making in libraries, Showers (2015) wrote that, for such residents, "part of you continues to exist when you are not online – your persona does not disappear, as it were, when you log off. This behaviour also means that we increasingly leave a trail – a 'data exhaust' – as we move across the web and interact with different spaces and networks" (p. 115). However, Showers' conclusion that "getting the data is, in some ways, not the problem. Rather, the problem is what we want to measure and why" (p. 115), completely ignores the ethical aspect of data-collection, surveillance of data bodies, and *measurement itself* in a culture of "hyper-control" (Stiegler, 2016, p. 58). Not only has the valorizing of quantitative measurement been criticized as one of the mainstays of neoliberal capitalism (Moore, 2018; Feldman & Sandoval, 2018), but advocates of student privacy maintain that libraries should be suspicious of their own abilities to collect student data (Goddard & Byrne, 2010; Jones & Salo, 2018). Libraries must also contend with unaccountable third-parties collecting and operationalizing data which libraries themselves may not (or at any rate should not) collect. The ethical implications of data-collection and metrics have been succinctly expressed in a Digital Library Federation explainer on the use of library data in research (Asher et al., 2018). The fact that such technological innovations *can* solve a real and long-standing problem in the library world (the problem of fragile and confusing access to licensed resources) while *at the same time* participating in surveillance capitalism has led to a "discursive struggle" (Hall, 1986, p. 40-41) within the profession which is far from over.

The concept of a data body moving through and leaving traces in disparate computer systems raises the question of what Maurizio Lazzarato has called "machinic subjugation" (Lazzarato, 2012) and Stiegler, "algorithmic governance" (Stiegler, 2016). We will look at these two concepts as they apply to the data body before proceeding to look more closely at the ways artificial intelligence technologies are used in CASA and RA21, along with the consequences this holds for privacy and the possibility of resistance.

Machinic Subjection and Algorithmic Governance

doi 10.20355/jcie29410

Left-criticism of technology, especially the "brilliant technologies" (Brynjolfson & McAfee, 2014) of the post-crisis moment, often focuses on the ways in which technology affects the formation of subjectivity, the ways in which the values, ideas, and personalities of individuals are formed by the fact that they are born into an "ecology" of machines (Negri, 2005, 87) and live their lives amid the hardware and software of post-industrial neoliberal capitalism. The idea of the intersection of technology and subjectivity within this left-criticism derives originally from Marx, who wrote about the subjective effects of the development of large scale systems of machinery in the section of the *Grundrisse* that has come to be known as the "fragment on machines." Marx argues that the development of technology involved a *qualitative* change from tools as "means of labour" or "fixed capital" to "an automatic system of machinery" (Marx, 1973, p. 692). Once this transition takes place, once machinery has been completely absorbed into the labour process, then tools of the labourer can be "set in motion by an automaton, a moving power that moves itself; this automation consisting of numerous mechanical and intellectual organs, so that the workers themselves are cast merely as its conscious linkages" (p. 692). The machine no longer appears as just another tool, whose function is to transmit the work of the labourer to the object; rather, the worker merely supervises and maintains the machinery as it operates itself. The end result is that

> Labour appears, rather, merely as a conscious organ, scattered among the individual living workers at numerous points of the mechanical system; subsumed under the total process of the machinery itself, as itself only a link of the system whose unity exists not in the living workers, but rather in the living (active) machinery, which confronts his individual, insignificant doings as a mighty organism. (p. 693)

Italian autonomist Marxists prematurely considered this process to have been achieved in the neoliberal turn to post-Fordism in the early-1970s. One such Marxist, the sociologist Maurizio Lazzarato, has since expanded on this view of the all-encompassing determination of the network of machines, focusing on the nexus of debt, technology, and subjectivity. For Lazzarato, indebtedness constitutes a parallel ecology to technology, and the subordination of all aspects of life to the maintenance of financial solvency at both the individual and the social level was, like the development of automation, a hallmark of the neoliberal shift. In the Fordist period, technological advances like the assembly-line were made to allow for the deskilling (and thus devaluing) of labour and the implementation of Taylorist efficiency. Now, in the post-Fordist period, technological innovation serves another social purpose: to outsource the keeping of promises (i.e. the repayment of debt) to the "objective" control of machines. In the orthodox view of post-crisis neoliberalism, in order to maintain the financial system and avoid further crises, all decision-making power should be placed in the hands of "unbiased" algorithms considered to be incapable of error. Critics of algorithmic technology, like Safiya Noble and Virginia Eubanks, have convincingly demonstrated that social values and prejudices are always embedded in seemingly objective algorithms (Noble, 2018; Eubanks, 2018). The pinnacle of "unbiased algorithm" perspective is the offloading of cognitive and evaluative processes onto "perfectible" machine learning and deep learning systems. The subjective effect of algorithmic decision-making is, however, to deprive individuals "of the possibility of evaluating risks and taking them; they are prohibited from challenging themselves in unexpected situations, working things out, and coming up with solutions. They are restricted to following the established

doi 10.20355/jcie29410

protocols and procedures" (Lazzarato, 2012, p. 143). In other words, machinic subjection seeks to drain all risk out of the socio-financial system, allowing for perfect forecasting, and therefore the mitigation of crisis. The political effect of this situation is "the fact that the process for evaluating and deciding is detached from any kind of democratic challenge or validation" (p. 143). In terms of library analytics, assessment, and student success, "the intent is to erase both anxiety and risk by finding assessment results that will nearly guarantee future employment so that present actions can be predetermined" (Nicholson, Pagowsky, & Seale, 2019, p. 67). At the level of everyday activity, Lazzarato uses the example of an ATM to describe the process of machinic subjection:

> The machinic functions without the "subject." When you use an ATM, it asks you to respond to the demands of the machine, which requires you to "enter your code," "choose your amount," or "take your bills"… There is no subject who *acts* here, but a "dividual" that *functions* in an "enslaved" way to the sociotechnical apparatus of the banking network. (p. 148)

Further echoing Marx's fragment on machines, Lazzarato concludes that "the credit card is like an apparatus in which the dividual functions like a cogwheel, a 'human' element that conforms to the 'non-human' elements of the sociotechnical machine constituted by the banking network" (p. 148). While this process has existed in neoliberal society since the 1970s, it has only recently begun to affect the academic sector and academic libraries, through the process Marx called "the subsumption of labour under capital" (Marx, 1976, p. 1019-1038), whereby capitalism expands into previously untouched areas via the "common sense" adoption of problem-solving and labour-saving technologies. Academic libraries – like universities themselves – had previously been more-or-less immune to the pressures of capitalist restructuring, but with the advent of the "knowledge economy," academic libraries now find themselves sites of commodity production and subject to the same logic as other capitalist enterprises (Popowich, 2018). As a result of this process of subsumption, the same dynamic of machinic subjection is applied to students who find themselves embedded in commodity relations they were not previously subject to. As a result, students are forced to become entrepreneurs, investing in their own human capital through education, as Foucault predicted (Foucault, 2008, p. 226). Such investment – like all financial activity – requires constant monitoring (in the form of metrics and analytics) and the elimination of risk through the gradual replacement of human decision-making by algorithms.

The first step in this process is to acquire the necessary data through a process of "primitive accumulation," which will make possible the machinic subjugation of the academic body. Such accumulation, as Silvia Federici reminds us, far from being consigned to an almost mythic past, continues to operate in the present (Federici, 2004, 12). We can understand this moment of the accumulation of student data as a corollary to the processes of subsumption and technological change in the onward march of capital accumulation and the subjection of the academic body through its proxy, the data body.

Just as primitive accumulation is an ongoing process rather than one consigned to a primordial past, for Bernard Stiegler, digital tracking technologies are simply the culmination of a process that has run throughout human history. This process, which he refers to as "grammatization"

(Stiegler, 2016, p. 29), involves the gradual breaking up of the undifferentiated flow of human experience into discrete pieces, beginning with the moment the human being learned to break up the flows of Paleolithic life into comprehensible parts. The invention of writing – the division of the flow of vocal sounds into discrete moments and visual representations – is the quintessential expression of this process. The age of big data or "algorithmic governance" is only the latest stage in a process in which individual behaviour and social relations are captured by computer interfaces (sensors, etc.), and "having become digital… in the form of binary numbers and hence as *calculable* data" then constitute "the base of an automatic society in which *every* dimension of life becomes a functional agent for an industrial economy that thereby becomes thoroughly *hyper-industrial*" (p. 19-20). The fundamental purpose of federated identity systems, whether or not they actually solve the problems of access and security, is precisely this colonization of all the dimensions of student life, in accordance with the broader technological imperatives of capitalist society. The concept of privacy, while important, seems at this stage inadequate to the issues at hand.

The move away from IP-based authentication is also part of the process of hyper-industrialization and the move toward algorithmic governance. Stiegler (2016) writes that through the implementations of data-driven networks, "processes of automated decision-making become functionally tied to drive-based automatisms, controlling consumer markets through the industry of traces and the economy of personal data" (p. 24).

The economy of personal data is what social media companies like Google and the vendors who form a large part of the RA21 steering committee (RA21, n.d.) are looking to unlock and control. The privacy and anonymity typically protected by libraries – now seen by capital as forms of risk – must be thoroughly undermined. In a passage that agrees in many respects with Lazzarato, Stiegler (2016) writes that the social network effect (of which "federated identity" is an important aspect) creates docile and manipulatable collectivities:

> In automatic society, those digital networks referred to as "social" channel these expressions [of individualization, like the individualism of Instagram "influencers"] by subordinating them to mandatory protocols, to which psychic individuals bend because they are drawn to do so through what is referred to as the "*network effect*," which, with the addition of *social networking*, becomes an *automated herd* effect, that is, one that is highly mimetic. It therefore amounts to a new form of *artificial crowd*, in the sense Freud gave to this expression. (p. 36)

What is at stake with federated identity systems that proxy the various student bodies for the benefit of capital is nothing less than the "automation of existences" (p. 19) themselves.

Creating the Student Data Body

The automation of a student's existence in systems of federated identity is markedly different to the traces left under IP-based or proxied authentication, due to the ubiquity of artificial intelligence techniques. Similar to the way in which "shadow profiles" for non-users are created in Facebook's social graph (Brandom, 2018), the various digital traces left within federated identity systems can be composed into rich or “thick” descriptions (Geertz, 1973, p. 9-10) of individual students made possible by the kinds of data used in both social networks and federated identity, known to the library world as “linked data.” Linked data is an implementation of a particular kind of artificial intelligence, symbolic or “good old fashioned” AI, which attempts to construct systems of knowledge based on atomic “symbolic descriptions” and the relations between them (Dreyfus, 1992, p. 18). Symbolic AI, in the form of linked data, underpins many of the newer networked technologies such as the Internet of Things and the ability of digital assistants like Siri or Alexa to search the web in response to a query. Fundamentally, linked data consists not of records, but of individual atomic statements which take the form <subject>-<relationship>-<object> (in linked data terminology, a relationship is known as the “predicate”). This fundamental data representation conforms to Minsky and Papert’s definition of a “symbolic description” as “a structure in which some features of a situation are represented by single (‘primitive’) symbols, and relations between those features are represented by other symbols” (Minsky & Papert, 1972). The combination subject-predicate-object into a symbolic description is called a *triple*, and one of the principles of linked data is that any agent (person or system) can make any statement (i.e. symbolically describe, create a triple) about anything else (Allemang and Handler, 2008, p. 7). The widespread deployment of linked data in, for example, social networks has created a situation in which statements about non-users (an offline friend mentioned by an online user in a post, for example) can easily be turned into a triple and stored in the social network of data. This triple – a simple data-point about a non-user – may be descriptively trivial in and of itself, but it costs nothing to store, and if and when another triple sharing either subject or object is created, a connection is instantaneously made, resulting in a thicker, richer description of the non-user than before. Over time, many such triples can be aggregated to produce a “triangulated” rich description of the non-user. The process is even simpler and more effective when applied to users actively producing content for the network itself.

The real power of this approach is not just in the aggregation of individual data traces but the ability of artificial intelligence techniques to “fill in the blanks” by making inferences based on the available data. In terms of symbolic artificial intelligence, new triples can be instantiated based on what must *logically* be deduced from two or more existing triples. A triple that exists in a linked data system is considered “true,” so when, for example, two true triples exist in a linked data system, a third true triple may be logically inferred from the truth of the first two. Take, for example, the following two triples:

> DR. PHUONG LE TEACHES ALL STUDENTS IN EDUCATION 501.
> JOHN SMITH IS ENROLLED IN EDUCATION 501.

A third triple,

> DR PHUONG LE TEACHES JOHN SMITH.

can be logically inferred by deduction even though such a data point exists nowhere in the aggregated data. Such inferences lay behind, for example, the identification of pregnant women from their harvested and data-mined purchase history (Duhigg, 2012).

The same process of aggregation and inference goes into creating an individual student's data body. Traces left in modern systems are represented as triples. At first, a given student is not necessarily identifiable as the subject or object of the triple. Over time, however, enough data is gathered together or inferred by users sharing more and more salient information that the identification of the student happens as a matter of course. When you add to the everyday collection of digital traces the necessity to login, either using a university's central user database or – in the case of CASA – Google itself, the identification of individual students takes place almost immediately. Subsequent data gathered on student behaviour happens as a matter of course and is, at least with CASA, not limited to activity taking place within a university's physical or virtual space, but Google's entire online ecosystem.

The fact that individual data traces – each innocuous in isolation – can be combined to create a thick description of individual students has two main consequences. First, the data involved can be dismissed as of limited danger or risk, for example as RA21 does in its "Security and Privacy Recommendations," published in July 2018:

> There are no significant risks which prevent [various projects] from moving forward. Residual risks from both a security and privacy perspective are LOW. The nature of the data involved is low value, i.e. not directly or easily attributable to any natural person, and appropriate safeguards are in place to mitigate confidentiality concerns. (RA21, 2018, p. 2)

This view of data traces in isolation ignores the second main consequence of the ability to easily (despite the RA21 claim) aggregate isolated data traces to produce a thick description: *emergence*. Looking at data traces in isolation mirrors the "methodological individualism" that characterizes positivist natural and social science, in which individual agents are considered ontologically and methodologically prior to aggregate social or communal structures (Bhaskar, 2015, p. 27-31). By taking an individualist perspective, social problems are thereby reduced to individual ones; in the case of identity management, problems at the aggregate level can be dismissed by focusing on the "low value" of individual data points supposed to be "not directly or easily attributable to any natural person." However, some non-positivist conceptions of science, such as the critical realism developed by Roy Bhaskar (1975, 1979, 1986) and others, argue that individual human beings who exist within networks of social and communal relationships (like the aggregate academic body) produce *emergent* effects which have a significance of their own. Emergent properties are those which are not possessed by any of the component parts of something (Bhaskar, 1986, p. 104), a common example being water, which possesses properties different from either hydrogen or oxygen on their own (Elder-Vass, 2010, p. 5). We can extend the idea of emergence to the properties of a given body, such as identifiability, which is not possessed by any of the individual, low-value data traces taken in isolation. For critical realists, the emergent structure, in our case the proxy data body, plays a role in the world different from the roles played by its constituent elements. Elder-Vass (2010) argues that the value of emergence is in its “potential to explain how an entity can have a causal impact on the world in its own right: a causal impact that is not just the sum of the impacts its parts would have if they were not organized into this kind of whole” (p. 5).

By theorizing the production of proxy student bodies out of isolated data traces from a left-perspective, we are committing to a conception of technology under capitalism as designed to further alienation, exploitation, and the production of surplus-value. The capture of individual behaviours under a regime of machinic subjection (Lazzarato) or algorithmic governance (Stiegler) therefore fits within a larger theory of technological innovation under capitalism, and it raises the question of resistance to such a regime. The concept of emergence was, for critical realists like Bhaskar, an integral element in his project of an emancipatory social science. In *Scientific Realism and Human Emancipation*, Bhaskar argued that a critical realist philosophy of science, one which recognizes the existence of unobservable but causally effective structures, can contribute to a project of human liberation precisely because such unobservable structures, defined by emergence, have great explanatory and demystifying power (Bhaskar, 1986, p. 103-104). The position taken by the RA21 “Security and Privacy Recommendations,” that individual data traces are so low-value that there is nothing to worry about, can be equated here with positivism in social science. Positivism in turn argues that only observable or detectable phenomena are or can be significant in the natural or social worlds, thereby excluding *a priori* the causal efficacy of emergent structures, supporting the mystification of the social world required by capital. Taken to an extreme, this process can lead to a situation in which complex phenomena are reduced to their atomic data points (Geertz, 1973, p. 17-18). Bhaskar’s critical realism, by taking emergent causal structures seriously, not only helps explain how such structures arise, but how they can then operate and have meaning within the social world. In this way emergence, exemplified in the proxied body composed of data traces, allows for an explanation that goes beyond the positivist position – which, as we have seen, denies any security or privacy concerns – to explain not only how the proxied body is constituted, but in what ways it can be an object of knowledge, surveillance, control, and profit within higher education. Such explanatory power arising out of the concept of emergence, Bhaskar argues, is a necessary component for any emancipatory project.

Conclusion

While it is the *emergent* data structure – the proxy academic body – rather than the data traces themselves that matters to students, teachers, librarians, administrators, and vendors in the federated identity space, emergence makes such structures sites of troubling ambiguities. To the student, for example, the individual data traces may seem harmless, nothing to hide, or a small price to pay for convenience, but the emergent data body quickly becomes the subject of targeted advertisement, ideological manipulation, mechanisms of grading and certification, surveillance, etc. Indeed, the very triviality of every individual data point is precisely *why* users often do not question exchanging privacy for convenience or functionality. For administrators, it is the data body rather than the individual data traces that are the object of success metrics, funding, support services, etc, even while each isolated trace (a grade held in a learning management system, for example) is carefully anonymized and subject to data-protection policies. For vendors, the data body is both the commodity and the worker: each user produces their own data traces out of their physical or virtual activity, but each data trace only becomes profitable as part of a larger social whole. Ambiguities such as these support and explain how, for example, "learning analytics privileges the individual over the social" (Oliphant & Brundin, 2019, p. 19), thus undermining any potential for social and collective activity or resistance.

By recognizing the role played by emergence in the creation of thick, rich data bodies out of trivial data points, or how widely-used artificial intelligence technologies "fill in the blanks" in such bodies, we can untangle the ways in which people are subjected to the machine ecology in which they live and work. Such recognition, and a commitment to challenge the "algorithmic governance" which serves to increase tracking, surveillance, risk-elimination, and the heightening of control, can help us to support and further the project of emancipation that underpins all critical theory. As technological advance increases the reach and scope of virtualized hyper-control, disciplines like education and librarianship, which have hitherto seen themselves as immune to such concerns, will need to adopt a critical, emancipatory approach to technology, data, and privacy in order to challenge and reverse the trends towards complete capitalist domination of education itself.

There are technological and policy approaches that can be brought to bear on some of these issues. However, initiatives like "privacy by design" (Cavoukian, 2010), similar to user-centred design or open-by-default data approaches, attempt to resolve the fundamental problems by conceiving of privacy (or user-experience, or openness) as something *left out* of existing design decisions rather than taking an explicit stance *against* privacy. This replicates a particular way of conceptualizing systems which sees a reified "privacy" as an isolated phenomenon that can be included or excluded from design decisions. Rather, questions of privacy cannot be isolated from the social relationships of which they form a part. Our attitude towards data, privacy, and systems reflects complex social structures and cannot be simply reduced to questions of bringing issues of privacy protection "in" at an early stage of the design process. Instead, the fundamental social relationship of institutions to users needs to be founded anew on less instrumentalist motives and designs. For example, instead of focusing on training, examinations, knowledge production, and even teaching, education could take a different approach, focusing on self-formation, self-cultivation, or *Bildung* (i.e. education as development), with "success" understood as a fundamentally subjective evaluation seen against a common cultural background. Hans-Georg Gadamer (2013) remarks that in *Bildung*, "that by which and through which one is formed becomes completely one's own" (p. 12). To reformulate education in this way not only changes the ways we think about student success, teacher effectiveness, surveillance, and privacy, but forces us to engage with the educational mission of our institutions and the role libraries can play in a new context. Of course, educational institutions are not autonomous but are themselves situated at the nexus of a complex of social and political relationships, which means that the transformation of educational approach suggested here is impossible without a refoundation of those relationships with a view to a more human, less instrumental social formation.

However, even prior to such a fundamental social change, there are things libraries can do. They can collectively leverage their bargaining power with third-party vendors, for example. We have seen this work in the case of non-disclosure agreements; it could work with surveillance, privacy, and data-tracking. This becomes more difficult when, for example, federated identity becomes part of the core functionality of a service (i.e. replacing IP-based access), but libraries are often reluctant to use their bargaining power to pursue normative goals, and this would be an opportune instance to act differently. However, as Dorothea Salo has been saying for a long time, the easiest thing would be to simply *not* collect user data, and certainly not to do so using tools which automatically store and report that data to third parties. "Libraries' default position," Salo has written, "is to collect data about their patrons. The correct default should be the opposite" (Salo & Kharfen, 2016, p. 60; see also Jones, 2019b). This approach, however, goes against librarianship's ingrained positivism and preference for evidence-based practice, and it would require us to be able to make decisions based on knowledge or understanding *other* than data.

This might, then, provide an opportunity for what we might call a *hermeneutic* turn in education and librarianship, in which interpretive methods, understanding, cultural context, and values trump data processing and "evidence" in decision-making and policy. Hermeneutics, for Gadamer, requires education as cultural development in order for the individual as a whole person to interpret and make decisions against a socio-cultural background. It would in theory be possible, then, to supplant the proxied data body with the real, embodied, cultural body of an institution's constituency; to work out our values, our policies, and our educational philosophies on a more human basis than we are currently doing. In any event, whatever approach is taken, it is clear that libraries and educational institutions cannot rely on corporations "doing the right thing." What is less clear is whether we can rely on our administrations to prioritize privacy and community values over logical practicality of data-driven policy. The coronavirus pandemic has provided a strong indication that whatever solution we adopt, it will need to be collective, democratic, and arising from the constituency rather than being imposed from above.

References

Allemang, D., & Handler, J. (2008). *Semantic web for the working ontologist: Effective modelling in OWL*. Burlington, MA: Morgan Kauffmann.

Anderson, R. (2019, April 16). *Researcher to reader (R2R) debate: Is Sci-Hub good or bad for scholarly communication?* Retrieved from https://scholarlykitchen.sspnet.org/2019/04/16/researcher-to-reader-r2r-debate-is-sci-hub-good-or-bad-for-scholarly-communication/

Asher, A., Briney, K., Gardner, G., Hinchliffe, L., Levernier, J. Nowviskie, B., Salo, D., & Shorish, Y. (2018). *Ethics in research use of library patron data: Glossary and explainer*. Digital Library Federation. https://doi.org/10.17605/OSF.IO/XFKZ6

Bales, S. (2015). *The dialectic of academic librarianship: A critical approach*. Litwin Books.

Bhaskar, R. (1975). *A realist theory of science*. Harvester Press.

Bhaskar, R. (1979). *The possibility of naturalism: A philosophical critique of the contemporary human sciences*. Harvester Press.

Bhaskar, R. (1986). *Scientific realism and human emancipation*. Verso.

Brandom, R. (2018, April 11). Shadow profiles are the biggest flaw in Facebook's privacy defence. *The Verge.* Retrieved from https://www.theverge.com/2018/4/11/17225482/facebook-shadow-profiles-zuckerberg-congress-data-privacy

Brynjolfson, E., & Macafee, D. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. Norton.

Cavoukian, A. (2010). *Privacy by design: The 7 foundational principles*. Office of the Information and Privacy Commissioner of Ontario.

Cheney-Lippold, J. (2017). *We are data: Algorithms and the making of ourselves.* NYU Press.

Deleuze, G. (1992). Postscript on the societies of control. *October, 59,* 3-7. https://www.jstor.org/stable/778828

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason.* MIT Press.

Duhigg, C. (2012, February 16). How companies learn your secrets. *New York Times.* Retrieved from https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Dyer-Witheford, N. (1999*). Cyber-Marx: Cycles and circuits of struggle in high-technology capitalism.* University of Illinois Press.

Dyer-Witheford, N. (2015). *Cyber-proletariat: Global labour in the digital vortex.* Toronto: Between the Lines.

Earle, S. (2016). The battle against geo-blocking: The consumer strikes back. *Richmond Journal of Global Law & Business, 15*(1), 1-20. Retrieved from https://scholarship.richmond.edu/global/vol15/iss1/2

Elder-Vass, D. (2010). *The causal power of social structures: emergence, structure and agency.* Cambridge University Press.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Federici, S. (2004). *Caliban and the witch: Women, the Body and Primitive Accumulation.* Autonomedia.

Feldman, Z., & M. Sandoval. (2018). Metric power and the academic self: neoliberalism, knowledge, and resistance in the British university. *Triple-C, 16(1)*, 214-233. https://doi.org/10.31269/triplec.v16i1.899

Foucault, M. (2008). *The birth of biopolitics: Lectures at the Collège de France, 1978-1979.* Palgrave Macmillan.

Fuchs, C., & Chandler, D. (2019). *Digital objects, digital subjects: Interdisciplinary perspectives on capitalism, labour and politics in the age of big data.* University of Westminster Press.

Gadamer, H-G. (2013). *Truth and method*. Bloomsbury.

Geertz, C. (1973). *The Interpretation of cultures: Selected essays*. Basic Books.

Goddard, L., & Byrne, G. (2010). The strongest link: Libraries and linked data. *D-Lib, 16*(11/12). https://doi.org/10.1045/november2010-byrne

Hall, S. (1986). The problem of ideology - Marxism without guarantees. *Journal of Communication Inquiry, 10(2),* 28-44. https://doi.org/10.1177/019685998601000203

Hartman-Caverly, S. (2019). Human nature is not a machine: On liberty, attention engineering, and learning analytics. *Library Trends 68(1)*, 24-53. https://doi.org/10.1353/lib.2019.0029

Hinchliffe, L.J. (2018, January 16). What will you do when they come for your proxy server? *Scholarly Kitchen*. Retrieved from https://scholarlykitchen.sspnet.org/2018/01/16/what-will-you-do-when-they-come-for-your-proxy-server-ra21/

Jones, K.M.L. (2019a). Introduction. *Library Trends 68(1)*, 1-4. https://doi.org/10.1353/lib.2019.0027

Jones, K.M.L. (2019b). "Just because you can doesn't mean you should": Practitioner perceptions of learning analytics. *portal: Libraries and the Academy 19(3),* 407-428. http://doi.org/10.1353/pla.2019.0025

Jones, K.M.L., & Salo, D. (2018). Learning analytics and the academic library: Professional ethics commitments at the crossroads. *College & Research Libraries 79*(3), 304-323. https://doi.org/10.5860/crl.79.3.304

Jones, K.M.L., Asher, A., Goben, A., Perry, M.R., Salo, D., Briney, K.A., and Robertshaw, M.B. (2020). "We're being tracked at all times": Student perspectives of their privacy in relation to learning analytics in higher education. *Journal of the Association for Information Science and Technology 2020*, 1-16. https://doi.org/10.1002/asi.24358

Jutting, C. (2016, August 2). Universities are tracking their students: Is it clever or creepy? *The Guardian*. Retrieved from https://www.theguardian.com/higher-education-network/2016/aug/03/learning-analytics-universities-data-track-students

Kilzer, R., Black, E.L., & Muir, J. (2008). Alternative solutions for off-campus authentication. *Code4Lib Journal, 3.* Retrieved from https://journal.code4lib.org/articles/73

Lazzarato, M. (2012). *The making of the indebted man: An essay on the neoliberal condition*. MIT Press.

Marx, K. (1973). *Grundrisse: Foundations of the critique of political economy.* Penguin Books.

Marx, K. (1976). *Capital, volume 1: A critique of political economy*. Pelican Books.

Minsky, M., & Papert, S. (1972). *Artificial intelligence: progress report.* MIT Press.

Montgomery, K. (2015, July 15). Proxy services are not safe: Try these alternatives. *Wired*. Retrieved from https://www.wired.com/2015/07/proxy-services-totally-unsecure-alternatives/

Moore, P.V. (2018). *The quantified self in precarity: Work, technology and what counts.* Routledge.

Moore, S. (2019, June 7). The politics of open access in action. *Samuelmoore.org*. Retrieved from https://www.samuelmoore.org/2019/06/07/the-politics-of-open-access-in-action/

Negri, A. (2005). *The politics of subversion: A manifesto for the twenty-first century.* Polity.

Newman, L.H. (2017, March 30). If you want a VPN to protect your privacy, start here. *Wired*. Retrieved from https://www.wired.com/2017/03/want-use-vpn-protect-privacy-start/

Nicholson, K., Pagowsky, N., & Seale, M. (2019). Just-in-time or just-in-case? Time, learning analytics, and the academic library. *Library Trends 68(1),* 54-75. https://doi.org/10.1353/lib.2019.0030

Noble, S.U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

OCLC. (2008, January 11). OCLC acquires EZProxy authentication and access software. *OCLC*. Retrieved from http://worldcat.org/arcviewer/2/OCC/2010/05/07/H1273247173434/viewer/file21.htm

Oliphant, T., & Brundin, M. (2019). Conflicting values: An exploration of the tensions between learning analytics and academic librarianship. *Library Trends 68(1)*, 5-23. https://doi.org/10.1353/lib.2019.0028

Popowich, S. (2017, February 1). The library systems disaster. *Redlibrarian.github.io.* Retrieved from https://redlibrarian.github.io/article/2017/02/01/library-systems-disaster.html

Popowich, S. (2018). Libraries, labour, capital: On formal and real subsumption. *Journal of Radical Librarianship, 4,* 5-19. https://journal.radicallibrarianship.org/index.php/journal/article/view/25

Popowich, S. (2019). *Confronting the democratic discourse of librarianship: A Marxist approach*. Library Juice Press.

Popowich, S. (2020). 'The power of knowledge, objectified': Immaterial labour, cognitive capitalism, and academic librarianship. *Library Trends 68(3),* 153-173. https://doi.org/10.1353/lib.2019.0035

RA21. (n.d.). About the team. https://ra21.org/index.php/about/

RA21. (2018). *WAYF Cloud and P3W Security & Privacy Recommendations*. Retrieved from https://ra21.org/index.php/results/ra21-security-privacy-final-report/

Rouvroy, A. (2013). The end(s) of critique: Data behaviourism versus due process. In M. Hildebrandt & K. De Vries (Eds.), *Privacy, due process and the computational turn: The philosophy of law meets the philosophy of technology* (143-167). Routledge.

Sadler, B., & Bourg, C. (2015). Feminism and the future of library discovery. *Code4Lib Journal 28,* 1-5. https://journal.code4lib.org/articles/10425

Salo, D. & Kharfen, S. (2016). Ain't nobody's business if I do (read serials). *The Serials Librarian 70*(1-4), 55-61. https://doi.org/10.1080/0361526X.2016.1141629

Schonfeld, R.C. (2018, January 22). Identity is everything. *Scholarly Kitchen*. Retrieved from https://scholarlykitchen.sspnet.org/2018/01/22/identity-everything/

Showers, B. (2015). *Library analytics and metrics: Using data to drive decisions and services.* Facet Publishing.

Sibbald, T. & Handford, V. (2017). Is there a metric to evaluate tenure? *Academic Matters, Winter 2017*. Retrieved from https://academicmatters.ca/is-there-a-metric-␟to-evaluate-tenure/

Simonite, T. (2019, October 17). The delicate ethics of using facial recognition in schools. *Wired*. Retrieved from https://wired.com/story/delicate-ethics-facial-recognition-schools

Srnicek, N. (2016). *Platform capitalism*. Cambridge, UK: Polity Press.

Stiegler, B. (2016). *Automatic society, volume 1: The future of work.* Cambridge, UK: Polity Press.

Tay, A. (2017, December 11). Understanding federated identity, RA21 and other authentication methods. *Musings About Librarianship*. Retrieved from http://musingsaboutlibrarianship.blogspot.com/2017/12/understanding-federated-identity-ra21.html

Tenopir, C. (2013). Building evidence of the value and impact of library and information services: methods, metrics, and ROI. *Evidence based library and information practice, 8(2)*, 270-274. https://doi.org/10.18438/B8VP59

Varnum, K. J. (2016). *Exploring discovery: The front door to your library's licensed and digitized content*. American Library Association.

Varnum, K.J. (Ed.) (2019). *New top technologies every librarian needs to know: A LITA guide*. American Library Association.

Webster, P. (2002). Remote patron validation: Posting a proxy server at the digital doorway. *Computers in Libraries 22(8),*18-23.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power.* London: Profile Books.

doi 10.20355/jcie29410