Notes on the Concept of Data Interoperability: Cases from an Ecology of AIDS Research Infrastructures

David Ribes
University of Washington
Human Centered Design and Engineering (HCDE)
Seattle, WA, USA
dribes@uw.edu

ABSTRACT

Data interoperation functions with the logic of a black box. Interoperation is achieved through front-loaded work and epistemically charged negotiation that thereafter become infrastructural, that is, supporting downstream actions without fully revealing the underpinnings that enable those actions. Drawing from ethnographic and archival investigations of data interoperation within and across an ecology of HIV/AIDS research infrastructures, this paper offers several sensitizing concepts for the investigation of how data are brought together and thereafter circulate. Data interoperability is historical, infrastructural, relatively irreversible, negotiated, epistemic, seamful and seamless, and is approaching the status of a general value rather than a specific means to an end.

Author Keywords
interoperability; data; infrastructure; ethnography; archival and historical research; HIV/AIDS

ACM Classification Keywords
J.4 Sociology
K.4.3 Computer supported cooperative work

INTRODUCTION

Reading the scholarly literatures focused on data one may be struck by a distinct disjuncture between those that study the work of data sharing, and those that study its consequences. Those who study *data preservation or sharing* will point to the immense difficulty, bottomless work, technical challenges and micro-political debates that accompany such efforts [4, 12]. They tell stories of situated data generation, of scientists that hoard their data for good or bad reasons, and information managers that struggle to keep up with constantly upgrading systems. Meanwhile, those who are concerned with the *consequences* of data interoperability are sounding the alarm, noting that our privacy is at stake, that new forms of data wrangling are revealing untold stories about our lives and society, offering new handholds for the manipulation of our activities, opinions and even emotions [30, 38].

A review of the broader field seems to reveal that with one hand we are emphasizing the immense labor and technical proficiency needed to achieve interoperability, while with the other hand we point to easy cross-contextual flows and the seeming ineradicability of data [29, 45]. We could summarize these contrary findings by counterposing the once felicitous countercultural phrase, now with increasingly problematized connotations, "information just wants to be free," with the findings of scholars of data sharing that can be summed up as "information just wants to stay still, keep quiet and degrade."

I engage the unarticulated intersections of these two sub-fields through a simple reformulation: interoperability is a fundamentally historical phenomenon. More precisely, data interoperation is a form of front-loaded practical work, negotiation and technical innovation that is thereafter black-boxed, largely forgotten, eventually taken for granted and naturalized as the inevitable technological trajectory for data. Only following interoperation do data flow with the ease promised by its advocates while evoking concerns about unanticipated or unintended uses. This is the topic of this paper; my goal is to offer an analytic avenue for bringing together these interlinked but currently balkanized topics by offering an historicist-methodological sensibility to the trajectories of data and their interoperability.

Rather than tackle this thorny topic at the societal level, something beyond the scope of this paper or my research, I will instead focus on a series of interlinked cases of data interoperation from my recent studies of an *ecology AIDS research infrastructures*. In this microcosm the benefits of data interoperation are clear – i.e., broader, longer or more heterogeneous views on HIV disease for social, biological or public health researchers – but these cases are also

CSCW '17, February 25 - March 01, 2017, Portland, OR, USA
ACM 978-1-4503-4335-0/17/03.
http://dx.doi.org/10.1145/2998181.2998344

marked with long and ongoing concerns about privacy and security, as the biomedical data of subjects must be protected even while continuing to circulate. Thus, this ecology is characterized by tensions analogous to broader debates about interoperability: the value of reassembling data for new purposes weighed against the dangers of unanticipated uses that threaten, for instance, privacy and security. I start with a case of data harmonization "within" an infrastructure long ago, and then two cases that track data "up and across" to another infrastructure: the first a story of data that circulate and the second where they do not.

Throughout I emphasize the situated, practical and deliberated quality of data interoperations. There is nothing natural, necessary or even tendentious about the interoperability of data. Interoperability is the outcome of sustained and arduous sociotechnical work and innovation over the last decades that only appear inevitable if one forgets the history of these activities.

I mean "historical" in two senses: firstly, that sociotechnical approaches to data interoperability have a long and winding past of successful innovations as well as dead ends [3, 13]. The histories of approaches to interoperability are not evanescent, they remain with us today: built into the very sociotechnical organization of data, interoperation does not disappear but carries forward to later uses of those data. Interoperability has been both a concern for information researchers in the past as well as being a lively field of research and innovation in the present, indicating that there is nothing "resolved" about the techniques for establishing and sustaining interoperability. In short, while some data are certainly interoperated, the subject of interoperability as a whole has not become historical in my second use of the term.

Secondly, I mean historical in the sense associated with the "closing of a black box" [24]: the process by that which is once the subject of debate or controversy, trial and error, and a great deal of effort, thereafter becomes backgrounded, taken for granted, and becomes a relatively stable resource, utility or commodity, in the sense that it is (more) easily, quickly, or even automatically accomplished. In short, interoperability, at its best, becomes infrastructural, supporting activity without revealing all that goes into that support [33], still subject to regimes of maintenance, repair and upgrade but only rarely of foregrounded debate [22].

Investigating the (still quite vast) microcosm of an ecology of AIDS infrastructures will not offer generalizable findings for understanding the societal investments and challenges of data. The technical and epistemic consequences of data integration are situated in the very activities of interoperation and thereafter black-boxing. Instead, I offer sensitizing concepts that will facilitate situated investigations of data, helping to make sense of the *historical, infrastructural, relatively irreversible, contentious and negotiated, epistemic, seamful and seamless* qualities of data interoperability (see table 1). I elaborate these concepts in relation to this ecology and then examine how interoperability, once a means to a specific end, may be shifting to a general value.

DATA INTEROPERABILITY

I will use the term data interoperability as an umbrella term for the constellation of concepts, approaches, techniques and technologies that seek to make heterogeneous data work with each other. There are dozens of such approaches and as many terminologies to match: commensuration, harmonization, integration, standardization, federation, mapping and innumerable others. I draw these together under a single term at the risk of erasing the marked differences in practical work, technical operations and downstream consequences entailed, though I will endeavor to remain specific when I discuss particular cases. The advantage of using the umbrella term is double. Firstly, to sensitize us to the broadly common goals across interoperability strategies: the smoother, easier or larger assembly of data. Secondly, the umbrella term serves to remind that a dataset, or the intersection of many datasets, are often the congealed outcomes of multiple interoperability strategies. Integrated datasets are historically inscribed by the repeated techniques of their interoperation – once again showing how "raw data" is an oxymoron [16].

An examination of the literature reveals that interoperability — whether data sharing, metadata documentation, commensuration or standardization — involves a great deal of practical work [40] that is often highly technical in both the technological sense of the word and the domain sense (i.e., scientifically technical), it is debated and sometimes even fraught with conflicts, and occasionally ends in failure (or at least retreating and regrouping) [3, 13]. The literature reveals a plethora of approaches and technological innovations to interoperability: schema mapping, ontologies, semantic integration, etc., along with multiple ongoing research agendas in the computer and information sciences to develop new, easier, faster or more robust sociotechnical approaches to interoperation of data [20, 21, 37]. Inspected over time, we see the rise and fall of particular interoperability strategies, but the outcome of a strategy (i.e., integrated data) remain with us as long as those data continue to circulate.

Interoperation is always partial, rather than either/or [31]. Data may be interoperated for one purpose or by one standard but insufficiently or incorrectly for another. For example, in the past I have investigated techniques of semantic interoperation of geoscience data [34]. The efforts of that group were to create an umbrella language for heterogeneous geoscience categories (i.e., a computational ontology). This set of categories could be used to describe data that had been generated under the diverse headings of the diverse earth sciences (e.g., geology, paleobotany, etc.). This "semantic interoperability" facilitated searches for

Interoperability is

Historically inscribed — Dependent on past work of generation and preservation of data, the available techniques and technologies for interoperation, and situated decisions about how to interoperate. Interoperated data travel only with traces of their trajectory to integration, if at all, while still carrying the consequences of that interoperation to future uses [24].

Infrastructural — Sharing many properties of infrastructure identified by Bowker and Star [5], interoperability facilitates or enables activities at the cost of rendering invisible (or murky) that which makes those activities possible.

Contested and negotiated –There are always many ways of integrating data. How data should be brought together is deliberated amongst actors doing that work, including: their goals, purposes, limitations and benefits, and the competing and evolving techniques of interoperation.

Epistemically consequential – What we can know, and how we know it, will be impacted by the decisions and methods that lead to data integration [34].

Relatively irreversible [6]– Interoperability displays qualities of path dependency or 'lock in' [9]. It may also, however, be subject to reappraisals in moments of contestation or re-interoperation, but this always requires an additional effort and if no traces of the trajectory to interoperation remain, then it may be irreversible.

Seamful *and* seamless – Working with data always involves practical and situated articulation work [41], but what specific practical work shifts following interoperation. Each successful interoperation (that is, use of data that have been interoperated without reopening the process that led to its interoperation) supports some seamless work, even if working with it is seamful [44].

Becoming a value [39], **norm [28], or virtue [8]** – Once a means to a specific end, interoperability is becoming an open-ended value. This is not an inherent property of interoperability but rather an historical outcome as sharing, openness, and documentation are encoded as regulative ideals.

Table 1: Sociotechnical qualities of data interoperability.

data, but still presented data in its unique heterogeneous databases. That is, the next step of bringing together the multiple datasets for analysis was left to those investigators: semantic integration facilitated discovery but not immediate use. As Bower and Star have noted of infrastructure more generally [5], interoperation is relative: what is interoperated by one standard or purpose may not suffice for another.

Evaluating whether a particular bringing together of data is right for the task at hand is a matter of assessing the data and inspecting the available documentation of that interoperation. Some data are interoperated with great care, capturing in metadata the transformations that have led to the final product [27]. But a great deal of interoperation occurs in a much less systematic manner, captured only in increasingly buried paper or digital archives, or not at all. Recently there have been efforts to encourage more systematic documentation of data transformations under rubrics such as 'reproducibility' [42]. Such efforts merit an investigation of their own, but in general they speak to a growing awareness of the topics I address in this paper.

Whether the traces of interoperation are readily accessible, a challenging work of archival spelunking or not available at all, once data are interoperated rediscovering how it was done and its consequences requires an additional effort. It is in this sense that I refer to interoperation as a black box. Whether a new dataset has been produced from past datasets or a query integrates data on the fly, once interoperated it is easier to rely on those data than to open the black box and understand their assemblage.

Tarleton Gillespie reminds us that digital logics operate differently than the traditional understanding of a black box with its stable inputs, outputs and an unchanging set of operations within [15]. If you leave your bicycle in the garage, and return to it the next day, it remains the same bike[1]. But the mechanisms of digital black boxes are different: when we return to a search engine it will appear much the same, as will the protocols for queries and responses (or APIs), but the search machineries and the data they are searching may have been subtly or drastically transformed by human, nonhuman or hybrid means. Such is also the case with interoperability, as deep within the machinery new forms of integration, commensuration or harmonization may be established, all the while appearing seamlessly to the unsuspecting querier of data.

METHODS & CASES

This paper is an ethnographic and archival-historical investigation of two "stacked" research infrastructures. They are stacked in the sense that the first case (the MACS) supplies the data that are interoperated by the second case (NA-ACCORD), in tandem with a broader ecology of AIDS research infrastructures. The NA-ACCORD integrates data from all while generating no subject data of its own (see figure 1).

An *ecology of research infrastructures* is a unit of analysis that recognizes that no infrastructure stands alone, but instead resides at the complex and evolving intersections of multiple sociotechnical organizations that broadly share goals, objects of study, standards and protocols, instrumentation, techniques and technologies, and collectively operate in common regimes of funding, policy and regulation. In this paper I am inspecting the ecology of

[1] A bike is a black box in the sense that the inspection of the artifact does not reveal the genealogy of technical innovations or competing social interests that led its contemporary form [32].

American biomedical cohort studies supporting HIV/AIDS science, funded by the National Institutes of Health (NIH) and regulated at the intersection of university IRBs, state and national agencies. Many members of the MACS and NA-ACCORD are scientists studying HIV disease in its various forms (e.g., as a virus, as a natural history, as associated with behaviors). However, inspecting data interoperability means focusing on those who work on information (a biostatistician, a data manager, a computer scientist) as they facilitate the integration and movement of data. Some of these are the same people.

My research team and I have inspected this ecology, first with a "deep dive" into the MACS and its history and thereafter by "following the data" [1] to those projects that seek to integrate them with those of other projects. We have participated ethnographically in current activities, such as all-hands meetings, or historically by inspecting internal documentation such as protocols or manuals of operation. Finally, we have interviewed members of these organizations, including scientists, technicians, participants and staff.

This paper is focused on the development of sensitizing concepts in the Symbolic Interactionist tradition [17]. Put briefly, sensitizing concepts tell the investigator where to look but not what to see, helping to guide empirical studies [18]. The paper focuses on aspects of the MACS or NA-ACCORD that help elaborate sensitizing concepts rather than seeking to fully characterize those projects; in the Grounded Theory tradition such targeted focus for conceptual elaboration is known as theoretical sampling [17]. I will refer to several other cohorts that NA-ACCORD seeks to integrate, such as 'clinical cohort studies' (a category explained below); the ecology of AIDS infrastructures is far vaster than the MACS and NA-ACCORD but in this paper I have theoretically sampled primarily from these two projects.

The MACS – The Multicenter AIDS Cohort Study

The MACS was founded in 1983 as an investigation of the natural history of AIDS. It is a longitudinal study of gay and bisexual men tracked over time as a prospective *cohort*. Investigators and staff generate their data every six months during "study visits" where they administer questionnaires (ranging in topic from social activities and locations, to toxic exposures and sexual behaviors) and medical interventions (measurements, such as blood pressure; or, specimen collection, such as blood, urine, semen, etc.) to the study participants. It is only a small selection of their data that they integrate with NA-ACCORD. The MACS does much more than produce subject data; in past CSCW [33] and in Science and Technology Studies (STS) [35] papers I have approached their assembly of instruments, specimen collections, sustained cohort and heterogeneous experts as the activities of a "kernel of a research infrastructure," but here I focus primarily on activities of data harmonization. Formally, this paper is an analysis across kernels, from the MACS and other cohorts to the NA-ACCORD.

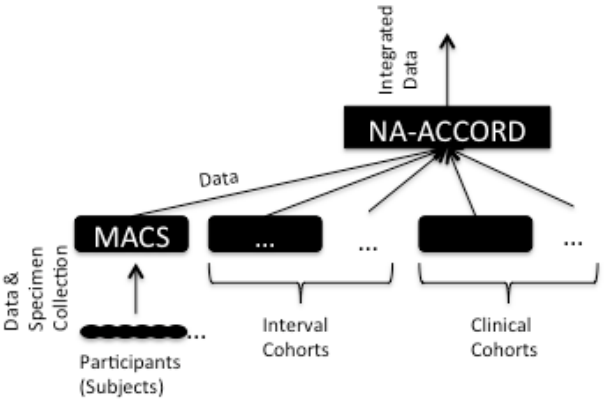

Figure 1: MACS and NA-ACCORD as "stacked" research infrastructures. The MACS and other cohort studies generate data that are integrated by NA-ACCORD.

NA-ACCORD – The North American AIDS Cohort Collaboration on Research and Design

Beginning in 2006, the NA-ACCORD was founded to integrate the data of over twenty cohort studies focusing of HIV disease [14]. Many of the contributing studies, such as the MACS, are long standing, some with decade long histories, each founded to examine a relatively narrow *cohort,* i.e., in the biomedical parlance, a group that shares some attributes and/or exposure risks, tracked over time, or longitudinally. Each cohort has its own complex organizational arrangements and study limitations, but the key promise of the NA-ACCORD is to facilitate aggregated and comparative studies across cohorts, an approach sometimes called 'supercohorts.' Rather than one-off studies, the NA-ACCORD seeks to enable faster, easier and more standardized cross-cohort investigations — in other words, the NA-ACCORD is a research infrastructure.

HARMONIZING SEROSTATUS DATA

I begin with a relatively simple but eye-opening example that I have previously recounted more extensively [35] of the heterogeneous definitions of "HIV positive" during the late 1980s. I will return to this little vignette throughout the paper as a stand-in for many of the key points I will make about interoperability: the value of harmonizing data, its potentially contested and historically shifting nature, its epistemological consequences and its forgotten and black boxed quality.

The MACS was founded in 1983, before the isolation of AIDS' causal agent in 1984: the Human Immunodeficiency Virus (HIV). Each of the two "co-discoverers" of HIV had also offered an assay for testing for the presence of HIV antibodies [19]. Those two approaches differed in key technical respects, searching for different markers of antibodies in the blood and with distinct limit conditions for evaluating whether a person was HIV positive, formally known as serostatus. The MACS of course keeps files about the serostatus of its cohort *as data*.

Following the isolation of HIV many pharmaceutical companies rushed to market biomedical testing kits that would serve to diagnose serostatus. A great deal was at stake, health and lives, yes, but also money, as these kits' first role would be to test the vast transfusion blood supply. Each of these kits drew on slightly different methods, materials and evaluation criteria for their results. Dozens of variations of these kits proliferated during the late 1980s, and many were versioned, i.e., pharmaceutical companies would release one and then another kit, adapting to emerging science.

Beginning in 1985, MACS scientists and doctors tested their cohort of nearly 5000 men for the virus using these kits. For their longitudinal study they sought to standardize findings, that is, their data, by using the same manufacturer of kits, establishing common protocols, and comparing findings. But as the science of HIV evolved, and as new HIV antibody tests were released, the MACS too had to rely on changing kits. In short, despite their best efforts, the MACS generated heterogeneous data about the serostatus of their cohort members.

Throughout this period, the broader HIV research community was engaged in a standards creation effort, seeking to define the appropriate markers to seek, the methods for testing, and the criteria for a positive reaction. It was in 1989 that a general consensus emerged that has since remained relatively stable. Consequently kits, and thus data, were also standardized.

For the MACS the years 1985-1989 left behind a set of heterogeneous data files, each relying on a distinct definition of HIV positive, specifically: what marker they sought out, and what criteria for reagent reactions to consider positive. This diversity was challenging for any scientist relying on these data, i.e., working with multiple competing definitions of HIV positive required arduous individual integrations of those data.

In order to remedy this MACS information managers and scientists engaged in a process of data "harmonization," deliberating the results of the various kits, their competing definitions of positive reagent reactions, and confirmation tests. A biostatistician recounted the development of a common definition that resulted in the creation of a unifying data file called "HIVDef89", or *the definition of HIV serostatus in 1989*:

> we worked with the investigators and the labs and came up with different algorithms for what is a seropositive, what is a seroconverter … [The result was the creation of] … a file that we call 'HIVDef89'.

Thereafter, rather than "redevelop who's HIV positive, who's HIV-negative, [or] when they seroconverted" a scientist could instead draw on a single harmonized dataset for their analyses. In the years that followed, this interoperated dataset has been reused hundreds of times for the investigation of innumerable objects of study.

I cannot speak at this level of granularity about many other datasets. This finding about HIV serostatus data took me a great deal of interviewing and archival spelunking. The reconstruction of a similar history for other data will likely be as challenging. And yet, in principle, it is how I approach data. It seems to me like a good initial premise that all datasets – whether scientific, social media or otherwise – have undergone various deliberated (or not) integrations leading to their tidy (or not) state at any later point in time. Only some of these transformations are well documented in the form of metadata that travels with the dataset; others are left in archival traces and memories (as with the tale I have told here); and for many, no documentation will remain at all.

This exemplar serves as stand in for my understanding that *data are always already interoperated*, that is, that a given datum, dataset, or their intersections with other datasets, have gone through processes that sought to harmonize heterogeneous forms of generation and storage through transformations that encode data in the same way. Scholars of research infrastructures and of longitudinal data have documented many cases of transformations in the ways that collecting "the same data" has changed over the years [5, 11, 23]. In each case some resolution across those changes is needed if the data are to continue to work together, e.g., comparatively, aggregated or longitudinally.

Here I have emphasized the harmonization of data produced over time, but even in this tiny vignette I have suggested several other ways that serostatus data were generated with an eye to interoperability, e.g., standardized kits, common protocols and cross-site testing. Interoperability is a defining quality of data, established and re-established from the moments of their generation. If a datum was not in some way interoperable with another, we would not consider it a datum at all.

In this case, the consequences of interoperation are vividly epistemological: how we know serostatus has immediate consequences for lives, yes, but for the MACS research agenda the epistemological consequences are in their downstream findings about the natural history of HIV disease. I return to this exemplar throughout the paper as I travel up- and down-stream to various integrations and data uses.

WHAT ARE HETEROGENOUS DATA AND WHY INTEGRATE THEM?

Before turning to the NA-ACCORD, which assembles the data of the MACS and many other cohort studies, I must first characterize, broadly, what it means to say that it integrates different kinds of data. I draw on a revealing design distinction between two kinds of cohort studies the NA-ACCORD brings together: interval and clinical cohorts. My goal is to understand what integration projects are "made of" by tracking back to the organizational machineries that generate those data. In this section I discuss the epistemic advantages and disadvantages of each, and finally, of growing importance, the vast cost differences for the two cohort forms. I describe these two types of investigative organizations in order address the advantages of integration by the NA-ACCORD in the next section, along with the dangers associated with black boxing that which is being integrating.

The MACS is an "interval" cohort study. This is a classical organizational form for biomedical investigations, usually in epidemiology i.e., the study of health and disease in populations. Well-known cohort studies such as the Framingham Heart Disease Study are interval studies. In such studies a cohort is recruited by the investigators and then tracked over time. Participants then return at regular "intervals" (thus the name): in the MACS the participating men return every six months to provide new data and specimens.

In contrast, clinical cohorts are built "on top of" existing medical provision. Their cohorts are formed by drawing together data from patients attending clinics. Sometimes referred to as a "virtual cohort," in these studies data are collected opportunistically as part of clinical patient care by integrating medical records, occasionally adding specialized instruments (i.e., questionnaires or assays), but largely relying on the existing instruments that are routinely implemented in medical care.

Clinical cohorts are, in part, already data integration projects, drawing together records and data from heterogeneous clinical environments. Clinical cohort studies are enabled by a past sociotechnical innovation in interoperability, in this case largely established during the 1990s. The challenges of integrating clinical data include, for example, distinct records at various clinics, forms written using varying terminologies, or working with the results of different versions of assays. In short, as with the harmonized serostatus data of the MACS discussed above, by the time the data of clinical cohorts reaches the NA-ACCORD their integration too is in the past.

Each of these two organizational forms for research generate systematic invisibilities and visibilities for desired and emergent objects of study, well documented in the biomedical literature [25]. What can be investigated or not depends on study design. I have room only to discuss one example: cohort make-up and the kind of generalizations that can follow.

Interval studies such as the MACS draw their cohorts from the general population unlike clinical cohorts that rely on populations receiving clinical care. An advantage is that interval studies like the MACS can recruit populations that do not have insurance or may not be receiving regular medical care. Thus, interval studies can make generalizable and comparative claims about insured and non-insured (or, under-insured) populations.

In contrast, most clinical cohorts render invisible the experience of disease for uninsured people. In the US not having insurance correlates with marginalized, stigmatized, poorer, and at-risk populations. Diseases will often manifest quite differently (not only 'worse') in different gender or racial groups, recreational drug users and so on, and a lack of insurance and treatment greatly impacts disease outcomes. Without such study populations, clinical cohorts cannot generate data on the variability of disease or effectiveness of treatment in those who are uninsured. This is an epistemic disadvantage of clinical cohorts[2].

However, when it comes to data, even in scientific circles, epistemic advantage is only one consideration amongst many. Cost is another. Even more so in recent years as biomedical funding allocations in the US have contracted.

Clinical cohorts have a vast advantage in costs. Rather than having to recruit and retain their cohorts, they enroll clinicians to do this work for them. Instead of hiring staff or technicians to conduct interviews and assays, clinical cohorts rely primarily on the results of tests that are already done. Thus, the costs of data collection are largely defrayed by relying on already operating routines of collection in clinics that are largely covered by the insurance of patients.

In contrast, the MACS and other interval cohort studies pay the cost for all the activities that lead to data generation. These studies "track" their cohorts by keeping extensive files on their participants and hiring staff to encourage and remind them to come back. They employ staff and technicians to administer interviews and collect specimens. They have labs (or contract out to labs) the various biomedical tests they require. Much of this "up front" data production work is conducted at the expense of a research funding award.

In general, and more recently in the face of challenged funding for biomedical research, clinical cohorts are advantageously placed relative to costs. Estimates vary widely but the annual cost per participant per year in an interval cohort is often cited as an order of magnitude larger than clinical cohorts. Cost becomes relevant as we examine the benefits of data integration in light of the tendency for interoperability projects to abstract away the genesis of its resources, which I address in the next sections.

The Advantages of Data Integration

The members of NA-ACCORD draw on and develop emerging capacities for heterogeneous data integration, large-scale computation and the (semi)automated tools such as medical record abstraction (i.e., drawing data from

[2] This is not a criticism. I am making a point about the differential value of these organizational forms for scientific investigations, or, in short, what scientific objects they can or cannot generate.

heterogeneous medical record file structures). Such integration projects have been gaining increasing momentum across the social, biomedical and natural sciences. The NA-ACCORD, founded in 2006, shares this ambition with virtually all "cyberinfrastructure" projects initiated around that time (e.g., in the geosciences [34], in the brain sciences [26], in ecology [27]).

Along with the promise that such infrastructures will make research easier, cheaper or faster, there is also the claim that they will make new kinds of knowledge possible. By pulling together data from many cohort studies, projects like NA-ACCORD open new avenues for investigation that may overcome the limitations of any individual cohort, and to enable larger 'statistically powered' studies that solidify more authoritative claims. Once again, the goals of interoperability are epistemological, promising to allow us to know more or differently.

Part of the advantages of supercohorts is that they can supersede the disadvantages of both interval and clinical cohorts. Projects like the NA-ACCORD draw together interval and clinical cohort studies and thus enable investigations using data from both. For example, studies relying on the integrated resources of the NA-ACCORD have compared HIV disease outcomes for those with and without insurance [36]. Data from interval cohorts can serve to represent the medical conditions and experiences of those without clinical care, filling in the gaps of clinical cohorts, and by combining the participants in many cohorts these studies can gain the 'statistical power' to make more significant claims than any cohort study on its own.

Recalling what data integration is "made of"

In this section I ask, by emphasizing integration are we at risk of forgetting what is being integrated? Projects like the NA-ACCORD are enabled by transformations in the information of its constituent cohorts, in particular the digitization and standardization of data that facilitate downstream integration. But what of the original cohort studies, and all the activities that lead to data that can be integrated? As Bowker and Star have noted of infrastructure more generally [5], interoperability projects tend to abstract away from the practical, technical and organizational activities that they rely on.

In interviewing the computer and information scientists working at NA-ACCORD I was surprised at the thinness of their knowledge of the MACS. Having spent the past few years digging into the MACS's long list of accomplishments its remarkable history was vivid to me: their contributions to fundamental knowledge of HIV, innovations in method, or the inspiring commitment of their participants during the horrors of the American AIDS crises in the '80s and early '90s. However, in retrospect the NA-ACCORD members' lack of historical knowledge of the MACS should come as no surprise. The MACS is only one of a score of projects that they integrate data from.

This is another example of the black boxing that occurs with interoperation: it creates an additional layer of abstraction from the practices of data collection, management, or past interoperabilities that have led to the constitution of the datasets that they will integrate. Each of those other cohorts, in addition to the MACS, too have complex, winding histories of tackling the tragedies of the AIDS epidemic and of innovative biomedical science; but those at NA-ACCORD know them only fleetingly. Nor do I for that matter, for those histories are not revealed in detail by studying just the NA-ACCORD.

Throughout the paper I seek to shift attention from data to the where those data come from, seeking to keep present what data interoperation tends to black box. Concretely, those include: the participants that volunteer in those studies; the doctors, staff technicians and complex instrument assemblies that generate data; and the distinct organizational forms called cohorts that enable certain research trajectories while foreclosing others.

When I speak to MACS investigators, very few know the story of HIV serostatus data harmonization I told above – long resolved and buried in its past – but almost all have a sense of the MACS' narrative in broad stroke: its 30 year history, its multiple scientific accomplishments, its dedicated cohort of participating gay and bisexual men. These understandings are learned as a feature of membership in the MACS' community of practice. But by the time the data of the MACS reach the NA-ACCORD all this is virtually lost, what remains is a valuable and reliable, if complex, longitudinal dataset, to be harmonized and integrated with a score of datasets from additional cohorts.

On its own, this is not a criticism. Rather, it is an assertion about the trajectory of interoperated data that tends to abstract away its practical histories. I have a related concern, and I return to the lower costs of data integration to make my point: in the increasingly economically pressured fields of biomedicine, less costly approaches such as clinical studies and (ironically) interoperation infrastructures are threatening the viability of more costly interval studies.

Evaluated on their own, supercohorts are even less costly per participant than clinical cohorts, and orders of magnitude less costly than interval studies. Demanding no subject data collection (often the most expensive feature of any cohort study), these projects bear only the costs of interoperation. Less expensive to fund, seemingly requiring far less effort on the part of investigators to collect and work with data, and more agile and responsive to emerging science, clinical and supercohorts cast interval cohorts as outdated organizational forms. A longer term, and institutional consequence of focusing on cost and the advantages of integration is that the innovations of clinical and supercohorts are producing a discourse that interval cohorts may be obsolete.

Today, the future of the MACS is uncertain, after over 30 years of operations it is on the cutting block. There are many reasons for this, too complex to discuss here (and, notably, it has weathered comparable challenges in the past); I stick to data interoperability and cost. Viewed as a matter of financial cost, "sunsetting" the MACS and other interval cohorts will free millions of dollars annually for the NIH to reinvest in other ventures. Perhaps this is the best course of action, I make no claims here; my only argument is that cost should be weighed alongside the epistemic advantages and disadvantages of these decisions: in cutting interval studies would we not be losing the key investigative advantages of such studies?

I have shown how interval cohorts have specific epistemic advantages over the other two forms. In fact, some of the advantages of supercohorts are fully reliant on the existence of interval cohorts, e.g., without interval cohorts like the MACS, supercohorts cannot make claims about uninsured populations; they just would not have those data. One NIH program officer described the NA-ACCORD to me, in confidence, as a "Ponzi scheme." Not implying anything elicit, that officer was referring its pyramidical structure, or what I called its "stacked" quality (see figure 1). The NA-ACCORD relies fully on its clinical and interval cohort studies for data to integrate; it contributes nothing back to those projects, instead contributing to the scientific enterprise by offering vaster assemblies of data. Viewed from an epistemic standpoint, without the MACS and other interval cohorts the NA-ACCORD will be diminished in its capacity to generate and investigate certain scientific objects. Because integration projects abstract away where its data come from, in collecting the "wheat" of cohorts' data they also render invisible the "chaff" that enabled its generation. Perhaps, the costs of data integration projects should be measured better not on their own, but in tandem with those data sources that enable their activities, a more holistic representation of what it costs to make their scientific claims.

In this section of the paper I have sought to keep the sources of data "on stage" at the same time as discussing data integration i.e., where those data and research materials come from, including the subjects/patients who provided them and the activities of data collection. I have sought to keep these present not (only) because of a humanistic interest in the contributions and work of these participants and scientists, but because forgetting these arrangements has consequences for the downstream reuse of their data, and for judgments about the value of these research infrastructures. The strengths and limits of the study designs that generate data are imparted to downstream efforts of integration. Integration may supersede certain limitations, but is ultimately still reliant on its multiple sources to do so. Interoperation tends to abstract away the practical and logistical trajectories that lead to that which is being in interoperated (data), but it does not escape the consequences of those trajectories.

HOW DO SUBJECT DATA TRAVEL?

This section will briefly outline the sociotechnical architectures for preserving the privacy of MACS participants. The personally identifying data that is held about MACS participants, whether HIV positive or negative, are never shared. This strategic non-interoperability has been established and reestablished repeatedly in the past thirty years. From its inception *the same project* that has sought to standardize and render its data reusable, has also sought to develop a sociotechnical system to ensure certain data never travel.

Clearly the identities of the MACS participants must remain confidential: it is a basic principle of modern human subject research. More than this, their identities are particularly sensitive, revealing of their sexualities (just knowing that a man was participating in the MACS would automatically identify them as gay or bisexual) and HIV status, during decades of appallingly prejudiced times (*perhaps* less so today, but still highly consequential). And yet, in order to track these men over time, *preserving* their identities is crucial for the MACS: both in order to find the men and encourage them to return for an additional study, and in order to create a dataset that longitudinally tracks the health trajectories of each man.

These data are kept separate from the rest (under "lock and key" for the first decades of the project, and now "behind password and encryption") and only a select group of staff and investigators have access to the men's real names, their addresses and contact information. Identities do not travel even so far as the coordinating center for the MACS: the centralized archives and those who work there nominally do not know the identities of participants. Identities do not make their way to the NA-ACCORD either: even the numerical confidential identifier that internally links MACS subjects across time is itself turned into a new identifier as data move to NA-ACCORD, adding an additional layer of protection as these data circulate.

MACS members preserve confidentiality at some expense, e.g., the financial and logistical cost of developing systems that keep personally identifying information in secured sites, but also at an epistemic cost their scientific enterprise: i.e., imagine the wealth of research that could be conducted if the MACS' vast troves of data could be linked to social media traces. In principle identifying data could be copied and reused indefinitely, never depleting their archives while generating granular views of disease trajectory, but in practice the MACS has evolved a complex privacy regime to ensure this is never the case. These efforts at sustaining confidentiality have been challenged repeatedly across their thirty-year history, facing technical innovations of re-identification, legal challenges from public health authorities, and a changing regulatory environment for personal identifying information. Across those challenges, MACS members have sought to preserve confidentiality.

I discuss this privacy regime here to counter any understanding that data "just wants to be free." Data should not be inspected "on their own" (for instance, from an information theoretic perspective), instead they should be approached relative to the sociotechnical systems that sustain them. Treated on their own, *as data like any other*, identifying information can be copied and shared as with any other information. But inspected as part of the ecology of infrastructures that sustain them, these data are best understood as part of an operation dedicated to ensuring they do not travel. This is a strategic non-interoperability, an intentional bulwark that limits the range of research so as to preserve confidentiality. Built into the organization of the MACS from inception – and regenerated as data are integrated by the NA-ACCORD – preserving this bulwark is easier (but not easy) to sustain than retrospectively attempting to establish a novel privacy regime.

DISCUSSION

In this section I return to the three examples above to draw out three key points about interoperability, and foreclose some possible misunderstandings about my claims (see Table.1). Data interoperability is *relatively irreversible*, and at others times completely irreversible. At its best, data interoperation can provide a *seamless* experience, but working with data is always locally *seamful*. And finally, I ask "is data interoperability a *value*?" and respond that, increasingly, it seems be becoming one.

Relative irreversibility

The first case I recounted above of harmonized serostatus data demonstrates a relative irreversibility. In the actor-network tradition Michel Callon has described irreversibility as "the extent to which it is subsequently impossible to go back to a point where that translation was only one amongst others" [6]. Here the translation is from one form of data to another, and a reversible translation would be an interoperation of data that could be conducted anew, in another way. The concept of relative irreversibility emphasizes that change is not an absolute impossibility, but that when continuity is "held together" by heterogeneous actors, change is a complex, multifaceted and daunting prospect.

The very archival methods I employed revealed that the original HIV testing data from the early '80s using heterogeneous kits remains available. So too do some documentation about those testing kits, and about the extensive and debated procedure for harmonizing the data. If needed, these data could (perhaps) be re-interoperated in a different way. But this would require a rather daunting effort; far more so than my historical digging to reconstruct this narrative it would require scientists and information managers to come together once again, recall and re-understand long forgotten technical details of testing, and come to a consensus for a new harmonization.

For integrated data where no traces of the original data remain or detailed metadata to regenerate them, a deep reconstruction of how they were interoperated is impossible. For those data that do carry with them extensive documentation about their transformations (an ambitious goal of the current "reproducibility" movement in science [42]), understanding their interoperation may be possible, but even then only at an additional cost in time and effort.

The original cases studied by Callon about electric cars in the 1980s are illustrative. Renault's efforts to build an electric car failed because of the interdependencies of cars with distributed fuel sources, technical limitations of batteries and electric motors at the time. Recent years have been demonstrating a slow reversal of the "lock in" of gas powered cars, but understanding the rise of the electric car demands looking well beyond the car to the creation of whole new networks of energy distribution, manufacturing and repair.

The interoperation of data, in the best cases, is not irreversible, but doing so requires a far greater additional effort than simply relying on extant assembled data. Most cases are not the best cases, and so, most historically interoperated data are irreversibly so.

Seamfullness & Seamlessness

I have emphasized the enabling properties of interoperation as it facilitates downstream work with data. For those interested in working with heterogeneous data, it is far easier to do so if someone has already done the labor of bringing them together. In ideal cases, drawing on such data may be seamless. However, as Janet Vertesi usefully reminds us through her studies of the international operations of NASA, no work is completely seamless. She draws on the concept of *seamfullness* as a cue for scholars of infrastructure that the common ideal of seamlessness is by no means total, and activities within and across such systems always require local articulation work [40].

In a facility in Spain, Vertesi uncovered a tangle of wires and nested cables [44]. Built to operate on the US electrical standard to support NASA equipment, the entire facility was an American enclave within Spanish electrical infrastructure. The tangle of wires contained a transformer that changed the electricity back to the Spanish standard. Tracing the cables from that nest she found the lamps and charging cell-phones of local Spaniards employed at the facility. The little arrangement of technologies she uncovered served to cross-connect multiple nested electrical infrastructures, a locally enacted interoperability.

Vertesi offers us a valuable object to think with, reminding scholars of the local work of making things work, and arresting visions of tidy or seamless operations that could wheedle their way into grand thinking about infrastructure. A corrective for anyone in danger of adopting a naïve understanding about the transparency of interoperation.

The case is similar with data. Today, researchers must always negotiate with their materials. Even the most well preserved, documented, and software supported analysis of data must be carefully worked over before, during and after analysis [43]. Working with data is seamful.

But seamfullness should not occlude the historical quality of interoperation that I have identified. The narratives of seamfullness and seamlessness are entangled [7]. Returning to Vertesi's example, pulling together three nested electrical systems relies on the established stability of each architecture and tools that have long enabled their interoperation: the stable 240v of the Spanish grid, the 110v of the American, and the little power-bars that have been available for decades to translate between these. The arrangements Vertesi uncovered are locally enacted and cleverly kluged, but those who do so are enabled by the historically established interoperabilities of electricity [10].

So too is working with data: past interoperabilities facilitate downstream articulation work. For example, today you can gain access the MACS public data set, available for a small fee on the National Technical Information Service website. It will arrive to you across networks of post on a CD as information encoded in the standard tables of the statistical software package R. To access them you may (seamfully) need dig up your CD drive, download and install R, but thereafter those data will be available to you to (seamfully) make sense of and use.

In those long tables you will find the HIV serostatus data for thousands of men over decades, but you will not see a scrap of the debates I described above about the changing definition of HIV positive; that interoperation is long resolved, black boxed, and instead (seamlessly) provided to you as clean rows of 1's and 2's. In short, once in hand, such data are seamless: they *are* HIV serostatus in the MACS regardless of the heterogeneous methods and criteria that led to them.

The full trajectory for any investigation relying on the data – i.e., getting data, and then working with them in an analysis – will always remain practically seamful even as the outcomes of past interoperations are presented in tidy seamless tables. Seamfullness and seamlessness are not opposites; they reside together in emergent, intercalated and interdependent relationships. Seamlessness supports seamful work, but no seamlessness works without a seamful resolution.

Is interoperability a value?

"Interoperability is not an end in itself" assert *Interop* authors Palfrey and Gasser [31]. At first glance, they are correct. We interoperate to achieve something else. But the matter is not so simple. In many, perhaps most, cases today the interoperation of data has no specific end. Rather there is an acknowledgement, or hope, that data may serve many future uses, such as in new investigations, to buttress the accountability of findings, or for public transparency. Openness (related to but distinct from interoperability) is increasingly encoded into American science policy: at the NIH most awards above 500K are required to share their data and include a plan to do so in their proposals; at the National Science Foundation long-standing requirements for data management plans are receiving increasing enforcement of late. I have focused here on a single interoperability infrastructure, but projects with comparable goals can be found across the sciences and beyond. In these cases, where the purpose of interoperability is open-ended, I tend to think that interoperability has become an end in itself, and arguably, a value [39], norm [28], or virtue [8].

In some sense, interoperability is a value I hold. I cannot question that the MACS and NA-ACCORD, with their carefully preserved and multiply repurposed archives, have contributed to our understanding of HIV disease and to the development and proper administration of treatments. Most of these studies and findings were not envisioned – could not be envisioned – at the inception of the MACS in 1983. Their data have had open-ended uses and that open-endedness has contributed immensely to the health of innumerable people. Ease of integration, comparability or aggregation may similarly benefit many other domains.

In this sense interoperability follows an archival logic [5]; a third meaning for "historical" in addition to the two I have emphasized thus far. As with preserving any archive, in part we interoperate for an indeterminate future: i) an archive contains more than we can know that it does; ii) it has value for people who have yet to use it; and, iii) future approaches and assemblies will yield findings not available to our methods today. As we have seen, interoperability is one strategy for stewarding data and renewing the value of the archive.

But there are also innumerable qualifications to an unhindered value for the archive, its integration with other archives and their emerging uses. I have room to discuss only one here, one that will stand-in for a broader consideration of unanticipated uses: informed consent. How could a man filling in forms in 1984 have conceived that those data would three decades later wind their way into a repository combining data from across North America and used to understand the effects of medical insurance? If his data has found this use, along with hundreds of other actual unanticipated uses, and an indefinite number of future uses, how can we speak of informed consent at all? This man's privacy has been protected, yes, but another core value of consent is that he will be informed, that is, have some control over the downstream use of his materials. As any scholar of IRB will point out, contemporary goals for data reuse and integration pose immense challenges to our current enactment of informed consent.

I raise the issue of consent as a stand-in for a much broader issue that it reveals: an assumption built into the valuation of interoperability that conflates future uses with good outcomes. There will be unforeseeable future purposes and uses for well-documented, easily accessible, interoperated data, but it does not follow that those uses will always be beneficial or ethically viable. A much subtler consequence is that since interoperation always relies on a set of situated decisions thereafter built into the interoperated dataset, if poorly documented that too will be black-boxed and thereafter consequential in unfolding ways.

In many circumstances today, data annotation, sharing and interoperation are a *means to an open-end*. In those cases we must keep in mind the thorny consideration that when finally put to a specific end, the downstream use of those data may be at odds with the intentions and commitments of those who contributed to the generation or circulation of those data. To say that someone or thing holds interoperability as a value, norm, or virtue means that they have committed [2], as belief or in practice, to the benefits of interoperability, that these will outweigh its dangers, and perhaps, that they are committed to tackling those dangers.

CONCLUSION

A favored maxim of historical research, particularly in the sociological vein of the study of science and technology, is that "it could have been otherwise." The phrase serves to mark the local, contingent and sometimes-serendipitous quality of historical trajectories that may otherwise seem inevitable. In this paper I have tried to articulate such a trajectory for data interoperability in an ecology of infrastructures. There are always many ways that data can be brought together; how it is actually done is a negotiation for the actors at hand, i.e., technical capacities, intended uses for data, competing interests, or epistemological commitments. Once brought together those data travel with only traces of their interoperation – often with none at all – largely leaving behind their conditions of assembly, but not their consequences. The interoperation of data can always have been otherwise, and thus so too its consequences: interoperated data and its downstream uses. Black boxed in the past, data can thereafter flow more easily, seamlessly and/or faster. Black boxing is neither an inherently negative nor positive quality, it is infrastructural, in the sense that it enables action at the cost of rendering invisible (or murky) the underpinnings of that action. Use of such seamless resources is never total; it is always accompanied by seamful articulation work.

Under the right conditions, data can be re-interoperated – i.e., assembled anew in a different way – but this requires an additional effort. When such an effort is particularly large we can call data interoperation relatively irreversible, when it is not possible at all, it is irreversible. In a final twist, in some cases (mostly, associated with architectures called "digital") the internal machineries of these black boxes may change over time with little or no marker of those changes for its users.

I began this paper with a discussion of the divide between two literatures that I feel should be in dialogue: those that examine the challenges of data production, sharing and preservation, and those that focus on the proliferation of data and problems associated with security and privacy. An historical approach to interoperability offers a bridge between these, as contemporary inflections of privacy and security concerns rest upon a long trajectory of data interoperation efforts.

Taken exclusively on their own, data follow the trajectory that scholars of data sharing and preservation have correctly articulated, in sum, as "information just wants to stay still, keep quiet and degrade." But following their contingent and locally negotiated positioning within networked systems of interoperation, data begin to take on the agency attributed in the phrase "information just wants to be free," along with its felicitous consequences of ease of access and reuse, and its more dangerous consequences for privacy, security, and other unintended uses. The maxim, "it could have been otherwise" gives us hope, it reminds us that we humans have a role in shaping the future trajectories of data interoperability, but the recognition of irreversibility points to the long backdrop of investments we have already made to the emerging technoscape of data, and the difficulties we will have in "making data flow otherwise," even if we do begin to make deliberated decisions today for how, what and when we would like to see data flow seamlessly.

ACKNOWLEDGEMENTS

I would like to thank the graduate students who have contributed to various facets of this research, and to the MACS and NA-ACCORD investigators and participants for access and support. This paper has benefitted from many conversations, long ago, with the Comparative Interoperability Project (CIP) team: Karen Baker, Geoffrey Bowker, and Florence Millerand. Lastly, I thank Janet Vertesi whose comments at a 2016 CSCW panel inspired the historicist arguments of this paper. This research has been supported by NSF/VOSS award #1322275 'Flexible Research Infrastructure' and the NSF/STS Scholar Award #1431219 'Technoscientific Flexibility of Research Infrastructure.'

REFERENCES

1. Susanne Bauer. 2014. From Administrative Infrastructure to Biomedical Resource: Danish Population Registries, the "Scandinavian Laboratory," and the "Epidemiologist's Dream". *Science in Context 27*, 02, 187-213.
2. Howard S. Becker. 1960. Notes on the concept of commitment. *The American Journal of Sociology 66*, 1 (July), 32-40.
3. Marc Berg. 2001. Implementing information systems in health care organizations: myths and challenges. *International journal of medical informatics 64*, 2–3 (12//), 143-156. DOI= http://dx.doi.org/http://dx.doi.org/10.1016/S1386-5056(01)00200-3.

4. Christine L Borgman. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology 63*, 6, 1059-1078.
5. Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting Things Out: Classification and its Consequences*. MIT Press, Cambridge, Mass.
6. Michel Callon. 1991. Techno-economic Networks and Irreversibility. In *A Sociology of Monsters? Essays on Power, Technology and Domination Sociological Review Monogrpah 38*, J. LAW Ed. Routledge, London, 132-161.
7. Matthew Chalmers and Areti Galani. 2004. Seamful interweaving: heterogeneity in the theory and design of interactive systems. In *Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques* ACM, 243-252.
8. Lorraine Daston and Peter Louis Galison. 1992. The Image of Objectivity. *Representations 40*, Fall, 81-128.
9. Paul David. 1986. Understanding the Economics of QWERTY: The Necessity of History. In *Economic History and the Modern Economist*, W.N. PARKER Ed., 30-49.
10. Paul David and Ann Bunn. 1988. The Economics of Gateway Technologies and Network Evolution: Lessons from Electricity Supply History. *Information Economics and Policy 3*, 165-202.
11. P.N. Edwards. 2010. *A Vast Machine: Computer models, Climate data, and the Politics of Global Warming*. The MIT Press, Cambridge, MA.
12. Paul N Edwards, Matthew S Mayernik, Archer L Batcheller, Geoffrey C Bowker, and Christine L Borgman. 2011. Science friction: Data, metadata, and collaboration. *Social Studies of Science 41*, 5, 667-690.
13. Gunnar Ellingsen and Eric Monteiro. 2006. Seamless integration: standardisation across multiple local settings. *Computer Supported Cooperative Work (CSCW) 15*, 5-6, 443-466.
14. Stephen J Gange, Mari M Kitahata, Michael S Saag, David R Bangsberg, Ronald J Bosch, John T Brooks, Liviana Calzavara, Steven G Deeks, Joseph J Eron, and Kelly A Gebo. 2007. Cohort profile: the North American AIDS cohort collaboration on research and design (NA-ACCORD). *International journal of epidemiology 36*, 2, 294-301.
15. T. Gillespie. 2011. Unpublished oral presentation at the Society for Social Studies of Science. Nov 11-13., Cleveland, OH.
16. Lisa Gitelman. 2013. *Raw data is an oxymoron*. MIT Press.
17. B.G. Glaser. 1978. *Theoretical sensitivity: advances in the methodology of grounded theory*. Sociology Press, Mill Valley, CA.
18. B.G. Glaser and Anselm Strauss. 1973. *The discovery of grounded theory: strategies for qualitative research*. Aldine Pub. Co., Chicago.
19. V.A Harden. 2012. *AIDS at 30: A History*. Potomac Books Inc, Dulles, VA.
20. Francis Harvey. 1999. Semantic interoperability: A Central issue for sharing geographic information. *The Annals of Regional Science 33*, 213-232.
21. C. Hine. 2006. Databases as scientific instruments and their role in the ordering of scientific work. *Social Studies of Science 36*, 2, 269-298.
22. Steven J Jackson. 2014. Rethinking repair. In *Media technologies*, T. GILLESPIE, P. BOCZKOWSKI and K. FOOT Eds. MIT Press, 221-240.
23. H. Karasti, K.S. Baker, and F. Millerand. 2010. Infrastructure time: long-term matters in collaborative development. *Computer Supported Cooperative Work (CSCW) 19*, 3, 377-415.
24. Bruno Latour. 1987. *Science in action: how to follow scientists and engineers through society*. Harvard University Press, Cambridge, Mass.
25. Bryan Lau, Stephen J. Gange, and Richard D. Moore. 2007. Interval and Clinical Cohort Studies: Epidemiological Issues. *AIDS Research and Human Retroviruses 23*, 6 (2007/06/01), 769-776. DOI= http://dx.doi.org/10.1089/aid.2006.0171.
26. Charlotte P. Lee, Paul Dourish, and Gloria Mark. 2006. The human infrastructure of cyberinfrastructure. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* ACM Press, 483-492.
27. F. Millerand, D. Ribes, K.S. Baker, and G. C. Bowker. 2013. Making an issue out of a standard: Storytelling practices in a scientific community *Science, Technology & Human Values 38*, 1, 7-43.
28. Michael Mulkay. 1976. Norms and Ideology in Modern Science. *Social Science Informations 15*, 637-656.
29. Helen Nissenbaum. 2009. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.
30. Leysia Palen and Paul Dourish. 2003. Unpacking privacy for a networked world. In *Proceedings of the SIGCHI conference on Human factors in computing systems* ACM, 129-136.
31. John Gorham Palfrey and Urs Gasser. 2012. *Interop: The promise and perils of highly interconnected systems*. Basic Books.
32. Trevor J. Pinch and Wiebe E. Bijker. 1984. The social construction of facts and artifacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science 14*, 1, 399-441.
33. David Ribes. 2014. The Kernel of a Research Infrastructure. In *Proceedings of the Computer Supported Cooperative Work (CSCW)* (2014), ACM, 574-587.

34. David Ribes and Geoffrey C. Bowker. 2009. Between meaning and machine: learning to represent the knowledge of communities. *Information and Organization 19*, 4, 199-217.
35. David Ribes and Jessica Beth Polk. 2015. Organizing for ontological change: The kernel of an AIDS research infrastructure. *Social Studies of Science 45*, 2 (January 8, 2015), 214-241. DOI= http://dx.doi.org/10.1177/0306312714558136.
36. Hasina Samji, Angela Cescon, Robert S Hogg, Sharada P Modur, Keri N Althoff, Kate Buchacz, Ann N Burchell, Mardge Cohen, Kelly A Gebo, and M John Gill. 2013. Closing the gap: increases in life expectancy among treated HIV-positive individuals in the United States and Canada. *PLoS ONE 8*, 12, e81355.
37. Amit P. Sheth. 1999. Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics. In *Interoperating Geographic Information Systems*, E. GOODCHILD, FEGEAS, KOTTMAN Ed. Kluwer Academic Publishers, Boston.
38. Katie Shilton. 2009. Four billion little brothers?: Privacy, mobile phones, and ubiquitous data collection. *Communications of the ACM 52*, 11, 48-53.
39. Katie Shilton, Jes A Koepfler, and Kenneth R Fleischmann. 2013. Charting sociotechnical dimensions of values for design research. *The information society 29*, 5, 259-271.
40. Carla Simone, Gloria Mark, and Dario Giubbilei. 1999. Interoperability as a means of articulation work. In *ACM SIGSOFT Software Engineering Notes* ACM, 39-48.
41. Carla Simone, Gloria Mark, and Dario Giubbilei. 1999. Interoperability as Means of Articulation Work. In *Proceeding of ACM Conference on Work Activities Coordination and Collaboration (WACC'99)* ACM Press, San Francisco.
42. Victoria Stodden, Peixuan Guo, and Zhaokun Ma. 2013. Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. *PLoS ONE 8*, 6, e67111.
43. Anissa Tanweer, Brittany Fiore-Gartland, and Cecilia Aragon. 2016. Impediment to insight to innovation: understanding data assemblages through the breakdown–repair process. *Information, Communication & Society 19*, 6, 736-752.
44. Janet Vertesi. 2014. Seamful spaces: Heterogeneous infrastructures in interaction. *Science, Technology & Human Values*, 0162243913516012.
45. Michael Zimmer. 2010. “But the data is already public”: on the ethics of research in Facebook. *Ethics and information technology 12*, 4, 313-325.