The kernel of a research infrastructure

David Ribes
Communication, Culture & Technology (CCT)
Georgetown University
3520 Prospect St. NW, Suite 311
Washington, DC, 20057
dr273@georgetown.edu

ABSTRACT

Infrastructure makes it easier, faster or possible for investigators to study objects of research. It does so by making available consistent and stable resources and services such as data, collaboration tools, sites of sample collection, or calibrated instruments. This paper offers the concept of the *kernel of a research infrastructure* as a new unit of analysis for investigating the enabling capacities of infrastructure. The kernel is the core resources and services an infrastructure makes available (called the cache), as well as the work, techniques and technologies that go into sustaining that availability (called addressing). By inspecting and comparing the kernel of two long-term scientific enterprises, this paper demonstrates how focusing on the kernel can help explain key qualities of research infrastructure such as flexibility and persistence in the face of dramatic changes to the objects, methods and practice of science.

Author Keywords

Infrastructure; long-term; sustainability; change; technoscientific flexibility; science; ethnography and archival research; grounded theory; kernel

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**] Sociology

K.4.3 [**Organizational Impacts**] Computer-supported cooperative work

General Terms

Management, Design, Human Factors, Theory

INTRODUCTION

Research infrastructures offer resources and services in the support of scientific and scholarly inquiries. They make it easier, or possible, for investigators to explore their objects of research. What particular resources and services are made available varies greatly by infrastructure. In studies *of* infrastructure, the most well investigated service has been facilitating communication and coordination [31], such as supporting geographically distributed collaboration or helping heterogeneous researchers work across disciplinary boundaries [26]. Also well investigated have been the archival functions of infrastructure, such as making it easier to curate scientific data [3], share them with colleagues [4] or enable reuse [12]. But infrastructures may offer other resources and services, less explored but equally important to the accomplishment of science, for example, analytic functions such as offering software tools [16] to support data visualization or making computing 'cycles' available [38]. Infrastructures may offer standardized instruments that enable the comparison of data across time and space [34]. They may play more cultural roles, such as offering researchers identity and affiliation, a venue for community formation, or a sense of vocation [17].

This paper offers the concept of *the kernel* as a new unit of analysis for the investigation of research infrastructure. The kernel is the core resources and services that an infrastructure makes available *and* the work, techniques and technologies that seek to sustain the availability of those resources over time. These two aspects of the kernel must always be taken together. A kernel approach examines resources and services as *entangled* with the work, techniques and technologies used to ensure their availability *as* resources and services.

For example, data is one such resource: some research infrastructures seek to make data sharing easier, for instance, in order to encourage reuse. Within such infrastructures there are actors (e.g., information managers) tasked with ensuring that those data remain available, well described, and in accessible repositories – in a kernel approach, this work is called *addressing*. The activities of sustaining availability of those data are entangled with the data themselves, for example, through the calibration of instruments [19], through data's redescription in metadata specifications [28], or by the ongoing maintenance work to keep data available online by web-service.

However, data are only one resource that infrastructure may seek to make available. This paper will focus on the set of core resources and services offered by particular infrastructures – in a kernel approach this is called *the cache*.

Focusing on the kernel, the cache and addressing will help explain many of the central capacities of infrastructure to enable and facilitate scientific research. By definition, a

CSCW'14, February 15 - 19 2014, Baltimore, MD, USA
Copyright 2014 ACM 978-1-4503-2540-0/14/02©$15.00.
http://dx.doi.org/10.1145/2531602.2531700

research infrastructure may have many purposes [41], however, one of the most explicit goals is the ability of infrastructure to support the investigation of new objects of research. *How do actors work to make the resources of infrastructure available for the investigation of new objects of interest?*

Focusing on the kernel will also help us understand the longevity of certain research infrastructures in the face of great transformations to the landscape of science. This is called *technoscientific flexibility*. Today, we consider flexibility to be a sought after virtue in the design of information systems. Flexibility may refer to, for example, the ability of a system to respond to changing user needs or to smoothly embrace emerging technologies [8]. But the ability to adapt to new scientific objects is a particular kind of capacity; for research infrastructure it is seminal, but one that has received very little scholarly attention. Focusing on the kernel will contribute to understanding technoscientific flexibility: *techniques, technologies and organizational innovations to adapt to changes in scientific objects, instruments and methods of investigation.*

To elaborate a kernel approach, this paper draws on an ongoing comparative study of two long-term research infrastructures: the MACS is the Multi-Center AIDS Cohort Study, and LTER is Long-Term Ecological Research. The empirical portions of this paper are dedicated to exploring the MACS and LTER kernels – which share surprising similarities given their distinct research concerns, while differing in revealing ways. The MACS is a *trim-poodle infrastructure* because it has a *single-core* kernel: its members work to standardize all their data and specimens and house them in common repositories. LTER is a *shaggy dog infrastructure* because it has a *multi-core* kernel: the diversity of their scientific materials and objects of study has posed great difficulties for standardization; today they have multiple repositories and standards for their data and materials.

The paper concludes with a discussion of how a kernel approach can serve to evaluate research infrastructure's ability to support the investigation of new objects of research.

What do we gain from a kernel approach? Understanding persistence, materiality and flexibility

The study of infrastructure has a strong intellectual lineage in CSCW [41]. A common focus has been on infrastructure that supports the doing of science, social science or humanities scholarship [31], often under the titles of collaboratory [13], cyberinfrastructure or eScience [39], but which in this paper will simply be called research infrastructure.

A great deal of research *on* infrastructure has focused on supporting collaboration of scientists [31]. A second focus has been to investigate how infrastructure supports data sharing and reuse [6, 42]. The two cases explored in this paper share these goals: the MACS and LTER are cross-disciplinary and seek to make their data available for new investigations. However, they also offer additional resources and services.

A kernel approach focuses on how infrastructures support the investigation of new objects of research, and to do so will broaden the analysis to include additional kinds of key resources and services – for example, the *materials* of science such as specimens and samples, along with where those materials come from. This approach places at center-stage renewal of the availability of resources and services: *supporting the ongoing investigation of new and old objects of research is the central goal of research infrastructure, and regenerating access to key resources and services is the central activity that enables infrastructure to do so.*

The Kernel

The driving metaphor in this paper is not about corn – a seed that grows or pops – it is a secondary metaphor inspired by software engineering. A kernel is the most important part of an operating system (OS), providing an abstraction layer that makes system resources available for applications.

Imagine the classic representation of the computing stack: interface and applications on the top, hardware on the bottom, and the operating system mediating these from between. From the perspective of the user, the kernel is essentially opaque, a black box that handles the operations "below." A user need not understand the machine language of device drivers, such as printers or scanners, or how to allocate activity in the central processing unit. An OS kernel is the infrastructure of practical computing, making the coordination of heterogeneous components appear seamless.

However, if we were to open this black box, as programmers that work on a kernel do, we would find millions of lines of code, built up and torn down over years or decades, sedimented by each additional contribution that sought to sustain the operations of a legacy device or enshrine a new computational capacity. It is from this perspective that this paper approaches the kernel – "from below," as an infrastructural inversion [5].

The kernel of a research infrastructure seeks to make certain resources that are needed in a scientific investigation available for use. It comprises two entangled aspects: the cache and addressing (see Figure 1).

The cache refers to the resources and services that an infrastructure makes available to support scientific investigations. In the cases explored in this paper, MACS and LTER, their caches have four features:

i. sites of ongoing data and specimen collection;
ii. an archive of data and specimens;
iii. standardized instruments, and;
iv. affiliated heterogeneous experts.

Other infrastructures offer altogether different configurations of resources and services. For example, GEON and LEAD [36], cyberinfrastructure for the earth and atmospheric sciences respectively, had no sites of collection, instruments, or data of their own. Instead their cache comprised data integration and visualization tools, i.e., they offered resources and services to interoperate data from heterogeneous databases and helped to make those data comprehensible by providing tools to render them in domain-specific visualizations. The Open Science Grid (OSG) offers only access to vast amounts of distributed processing power and storage [38]. We will return to these examples in the discussion.

The question: "What resources and services?" has as many answers as there are infrastructures – it is an empirical question that can only be answered in the particular. The composition of the cache is specific to the history of an infrastructure.

Over time, new kinds of resources and services that an infrastructure may offer are innovated – such as data integration or cycle sharing – but the "classic resources" remain just as important to doing science. The emphasis in recent studies of infrastructure development efforts has been on digital resources, such as data. However, within science, materials, such as specimens and samples, are crucial for inquiry. For example, one of the most valuable resources that the MACS offers its members is a specimen archive of blood that goes back nearly thirty years. The LTER site in Baltimore offers something similar yet distinct: an archive of stream water samples that stretches back thirteen years. To illustrate the kernel, we will return to the examples of blood and water throughout the paper.

The key point in examining a cache is to expand the unit of analysis to include all the core resources and services an infrastructure makes available, eschewing any fixation on 'high tech' resources in favor of understanding how features of the cache – new and old, material and computational, digital and paper based – are rendered available for the investigation of a research object.

The second aspect of the kernel is called addressing, which refers to the work, techniques and technologies that seek to ensure that resources and services are available for use, in a stable manner, and over time. As with the cache, addressing is highly heterogeneous and specific. Returning to blood and water, we will find that such scientific materials are not distinct from the information systems of infrastructure: an aliquot of blood is only meaningful if it continues to be tied to the individual who donated it, their blood pressure, their weight, and all the other blood samples they have donated over the years. Water must be tied to the stream that it came from and the date of its collection. All of these data are recorded in tandem with collection of blood and water: an information management issue. Within the MACS and LTER there are actors who work to ensure that those samples are collected in comparable ways across the years; that they are labeled and stored in ways that will preserve their useful material qualities; and that they can be retrieved for later use with relative ease. All of these, and many more, are the activities of addressing.

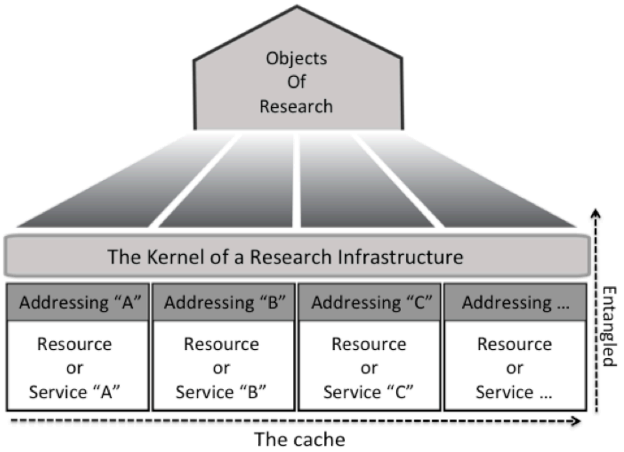

Figure 1: The kernel of a research infrastructure. Resources and services are made available for the investigation of objects of research through ongoing activities of addressing. ***Image credit: Jake Fagan***

The term "entangled" is essential to understanding the relationship between the cache and addressing. To clarify what is meant by entanglement, an example from traditional infrastructure is revealing: we may initially think that drinking water and the pipes that deliver it to our home should be thought of as discrete, however this is misleading. Part of the infrastructure of drinking water includes filtering silt, extracting pollutants, inspecting for bacterial cultures, and adding chlorine. *Infrastructure transforms that which it makes available.* If anything other than water in this form were to pour from our taps, we would consider the infrastructure to be broken, dangerous, a matter of public safety and welfare. Similarly, a research infrastructure operates upon its resources: data must be cleaned, standardized and annotated; instruments calibrated across geographic sites and over time; specimens preserved in the same manner, properly stored and labeled [37]. The term entangled will serve as a shorthand to remind us that in the approach outlined in this paper we should not distinguish the resources and services of a research infrastructure from the work of regenerating their availability[1].

The final aspect of the kernel is that it facilitates change to itself. That is, many actors within infrastructure are reflexively engaged in a process of monitoring the kernel and the research activities it is intended to support. As

[1] The concept of entanglement is a departure from the kernel OS metaphor. In its software sense, a kernel is distinct from the resources that it makes available, i.e., the operating system is not the printer, memory or CPU. However this makes little sense for a kernel approach to infrastructure. Notably, the term 'entangled' is not native to the vocabulary of OS kernels, it is an intentionally mixed metaphor so as to better remind of this break.

research goals shift over time, the kernel itself must change to support the investigation of new objects. For example, over time new disciplinary specialists have joined LTER: as ecology has become more and more concerned with human ecologies (e.g., a city), LTER has added specialists concerned with human activity: social scientists. Social scientists have different instruments and different kinds of data (e.g., surveys) and consequently the LTER kernel has been adapted to help facilitate these kinds of inquiries and preservation of these kinds of data.

The kernel is built with a prospective expectation of change and there are procedures and technologies within the operations of infrastructure that facilitate such changes to the kernel – these are examined below as *elaboration* and *extension*.

Paul Edwards, in his studies of another large-scale and long-term research infrastructure – the vast machine of climate science – has noted that "infrastructure" is a particularly useful concept for understanding flexibility in science:

> The concept of knowledge infrastructure [see 10] resembles the venerable notion of scientific paradigms, but it reaches well beyond that, capturing the continuity of modern science, which keeps on functioning as a production system even while particular theories, instruments, and models rise and fall within it. [.] I prefer the language of infrastructure, because it brings home fundamental qualities of endurance, reliability, and the taken-for-grantedness of a technical and institutional base supporting everyday work and action. [9]

The concept of the kernel, repurposing of its features such as data and experts, and change to the cache through elaboration and extension, are contributions to understanding *technoscientific flexibility*: how infrastructure prepares, manages and responds to change in science even as its objects of research, methods of investigation, and the experts who enact that research are transformed.

CASES AND METHOD

The design of this study is historical-comparative and ethnographic-comparative. Comparisons are of two case studies of infrastructure supporting scientific research, founded approximately thirty years ago, and continuing in the present.

The MACS – The Multicenter AIDS Cohort Study

The MACS was founded in 1983 to investigate the causes and modes of transmission of a very poorly understood disease: Acquired Immunodeficiency Syndrome (AIDS). To do so, the MACS was established as a multicenter study, a distributed biomedical organization in four American cities: Los Angeles, Pittsburgh, Baltimore and Chicago. In these cities they recruited thousands of gay or bisexual men (the most visible 'at risk' population), and tracked them over time – as a prospective cohort – by administering biannual 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.).

LTER – Long-Term Ecological Research

LTER was founded in 1980 to "understand general ecological phenomena which occur over longer temporal and spatial scales" [27]. The sense within the scientific community was that to study ecological change, which occurs over decades or centuries, the investigation itself should be temporally scaled to match. To do so, LTER established six sites for long-term data and sample collection. Over time LTER has added many new sites and has shut down others; in 2013 it has twenty-six sites of investigation. These sites span a great array of terrestrial and aquatic ecosystems such as deserts, prairies, lakes, and estuaries. For example, the Hubbard Brook site is focused on forests and small streams. There, they have tracked the temperature of streams for over thirty years; they also sample those streams by bottling water and silt.

Method

This paper is focused on theory exposition in the Grounded Theory tradition [7]. The paper targets (or *theoretically samples* [14]) the empirical features of the MACS and LTER most relevant to understanding the cache, addressing and change to the kernel.

In order to track change to science in these research infrastructures, my research team inspected the corpus of associated publications within each infrastructure, year-by-year over each thirty-year history e.g., scientific publications, funding proposals, reports, requests for proposals (RFP), and other documentary sources including archived websites.

Members of my research team participated ethnographically in current activities, such as data collection outings, All-Hands meetings, executive and advisory meetings. I have worked closely with six sites of LTER over nine years, have been an observer of the MACS for two years as well as participating in the design of a study for its Baltimore site; one of my research assistants worked for the MACS for one year.

To better understand information management practices, I theoretically sampled portions of data archives in the MACS and LTER. Finally, I have interviewed members of these organizations, including scientists, technicians and staff; in the MACS this has included interviews and focus groups with participants (subjects) in the study from the Baltimore/Washington site.

THE MACS AND LTER KERNELS

For two enterprises dedicated to studying very different objects, the MACS and LTER have surprisingly similar kernels. Both organizations have sought to persistently make available four sets of resources and services. This section outlines each feature of the cache in turn: sites, archives, instruments and experts. The next section will explore how each feature has enabled the MACS and LTER to continue supporting science in the face of fundamental transformations to their objects of research.

The cache: Resources and services of MACS and LTER

i. Ongoing sites of data and sample collection

When we think of doing science we often think of data, but in long-term collection enterprises such as the MACS and LTER, *the source* of those materials is nearly as important as the data and specimens themselves. The term 'site' refers to the source of collection for scientific materials. The heart of a prospective longitudinal study is the ability to ongoingly produce novel data and samples, i.e., when one cannot answer a research question using extant data and samples, ongoing sites enable the collection of new materials to do so.

In the MACS, the sites of collection are the cohorts of gay and bisexual men who have, for the last thirty years, every six months, travelled to clinics in their respective cities to once again fill out questionnaires and donate the blood and tissues that then become specimens. All data and specimens come from these men and two further cohorts that have been recruited since: nearly 5000 from the first cohort recruited in 1984-5, with a total of nearly 7000 across all cohorts. The MACS continues to track the men who remain in the study to this day.

In LTER the sites of collection – and what is collected – are far more diverse. In 1980, LTER was founded with only six sites, today there are twenty-six. Each site is distinct and the focus of different kinds of research; members collect many different kinds of data and samples. An example that will serve as a touchstone throughout the paper is drawn from the Baltimore Ecosystem Study (BES) site of LTER where every week a team of technicians and graduate students head out in a van to circle the Baltimore area, stopping at twelve spots to measure stream temperatures, to check local rainfall levels and to collect four bottles of stream water.

ii. Archives of data and specimens

The MACS and LTER have a mandate to keep data and specimens available for the investigation of new and old objects of research. Longitudinal datasets and collections are key contributions of both enterprises, allowing their scientists to explore their most ambitious and central objects: the natural history of HIV/AIDS (i.e., how the disease progresses and varies in individuals over time), and ecological change that occurs over decades.

When the participating men in the MACS come to clinics they enact a routine called "the visit" in which they fill out questionnaires on their well-being, sexual behaviors, drug use, medications, and many others. For the core questionnaires that have been administered since 1984, to this day, the men continue to fill out bubbles on "scantron forms" using a #2 pencil. During a visit, blood is always taken; other materials have come and gone. Depending on when in the thirty-year history of the MACS we were to join the visit, we could observe the collection of other materials such as urine, fecal swabs, or throat gargles. In 2012, the MACS reported the storage of 1,638,409 aliquots of urine, cells, plasma and serum.

Samples are not simply a step in the translation of a site into data [22]; they have value unto themselves. Just as with data, they are preserved and curated. In LTER/BES two of every four bottles of water that leave each streambed in Baltimore are ported to Millbrook, New York. There they are placed in cold storage with thousands of other bottles, neatly identified with glued-on labels that mark their site and time of collection. Samples are not discrete from data, they too are entangled, i.e., a sample of water is only scientifically meaningful if it is tied (with data), to a specific place of collection, a specific time, and all the rest of the water and other data collected from that streambed over the years.

These scientific data and materials circulate in many ways across their respective scientific communities. Today, the most well documented data are available online for immediate download; other datasets are available upon request; and still others exist only as paper documents in filed archives. Available in principle, such data may be tedious or impossible to access in practice. Similarly, scientists can gain access to material samples. Ecologists travel directly to the sites of cold storage to analyze stream samples, or, a styrofoam-packed and frozen aliquot of blood can be delivered to biomedical researchers by courier.

iii. Standardized instruments

Instruments (and their associated practices) are the devices that transform sites, materials and subjects into data – they are key information technologies in the doing of science. Both the MACS and LTER are geographically distributed organizations with the goal of longitudinal synthesis and comparison. Comparing data and samples across time and space (usually) means collecting them and analyzing them in comparable ways, and this requires the ongoing standardization of instruments.

For example, a key instrument for the MACS has been the flow cytometer. This largely automated machine runs a laser over flowing tissues, such as blood, in order to classify and count entities like CD4 white-blood cells [19]. Previous to the availability of the HIV test in 1985, measuring CD4 counts (then called T-helper cells) was one of the only methods for detecting AIDS before the onset of symptoms.

At the founding of the MACS, the principal investigators decided to buy the same kind of flow cytometers from the same vendors, using the same reagents to mark cells, and to use the same brand of calibration beads to standardize them. Even this is not enough: on a regular basis, the very same specimen of blood is circulated across the four sites of the MACS to compare measurements for each instrument. Without this work of ongoing standardization, comparing across sites would be methodologically challenging or impossible.

In LTER/BES, they have considered automating the collection of stream samples. They sought to delegate the weekly work of bottling water to machines by installing a device at each site that 'sips' the stream at regular intervals. However, such a transformation in the sampling method threatened the comparability of the archive: water collected in a new manner could shift key measures in ways that biased the data. To this day in Baltimore they run the automated machines *and* continue to slog into the middle of the stream to collect bottled samples by hand. Now two sets of data, one goes back a decade longer than the other, preserving comparability at the expense of more labor.

iv. Heterogeneous experts

HIV disease and ecology are complex; they do not sit tidily within the confines of disciplines. To support ongoing investigation, the MACS and LTER maintain a network of heterogeneous experts. The MACS' founding members were epidemiologists, virologists, and research doctors. But illness is as a much a sociological object of research as a biological and medical one. In its thirty years the MACS has increasingly extended its base of expertise to sociologists, psychologists and many others.

Experts are not only scientists. LTER has an information manager at each of its twenty-six sites. These specialists are specifically tasked with curating an ongoing archive, facilitating data sharing, and prospectively planning for a technologically evolving datascape, i.e., from flat-files and floppy disks in 1980, to relational databases, web-services and metadata specifications in 2013.

Facilitating communication between its geographically distributed members is also a service the MACS and LTER offer, often through relatively "low-tech" approaches such as member directories, but more recently through specialized discovery tools. Both organizations host face-to-face gatherings, e.g., "All-Hands" meetings that assemble the majority of their members, or more targeted get-togethers by topic or goal. Such meetings are opportunities to share findings, methods and common challenges, but also, to get to know each other *as* MACS and LTER members, rather than as specialists at disciplinary conferences.

CHANGING OBJECTS OF RESEARCH

Change to the very root of the scientific enterprise can be vividly illustrated through both the MACS and LTER. This section outlines changes in broad strokes. The next section inspects these changes more granularly to explore how the MACS and LTER caches helped them adapt to these changes.

When the MACS was founded in 1983, no causal agent for AIDS was known. A central goal of the MACS was to participate in the search for AIDS' etiology. It was a year later, in 1984, that French and American researchers identified the Human Immunodeficiency Virus as that agent, and it was not until 1985 that a commercial assay became available to test the thousands of men in the study for that retrovirus.

I will not further recount this story, which is described in greater detail elsewhere [40], but instead emphasize the technoscientific change: What the MACS was tasked to study in 1983, its object of research, was a newly recognized but unknown and ultimately deadly disease, AIDS. At that time many causes were posited – an infectious agent, yes, but environmental and behavioral causes were also being explored. The initial purpose of the MACS was to help pin down a cause. That mission vanished with the discovery of HIV. This is quite literally an ontological transformation: in 1983 we had no recognized such entity in the world, and in 1985 we had both HIV and people who were seropositive. Thereafter the MACS became a study of the natural history of HIV.

In LTER scientific change has been less immediate and life-and-death, but ultimately, one that is proving to be part of a more radical cosmological shift. LTER was founded as a study of six sites, each a distinct biome: a large, naturally occurring collection of flora and fauna. In 1980, the framing was largely that these biomes would be studied discretely and comparatively, as unique biomes. Since then the science in LTER has gone through two major refigurings. Firstly, it is now a much more globalized enterprise; biomes are still units of comparison, but they are also modeled as part of an interconnected global system. LTER collaborates and coordinates with "other LTERs" all over the world through its international coordinating body: ILTER. Secondly, humans are increasingly framed as part of ecology rather than an experimental variable to be controlled ("disturbances" as they were called in 1980 [29]). The first six biomes were chosen largely for their purity: "areas of relatively pristine, preserved ecosystems" [27]; they were "nature on its own," devoid of human intervention. Today, in tandem with changes in climate science, humans are treated as important actors in reshaping *and* sustaining ecological biomes. In short, the vision of LTER today is more globalized, and the role of humans to ecology is more central.

Change to the kernel: Repurposing, elaborating & extending

The narratives above are a view of scientific change from "1000 kilometers up." Each of these broader changes was actualized as innumerable more granular reorientations in objects of research. This section will focus on how each kernel was *repurposed* for the investigation of these new objects, and then examine two forms of punctuated change to the kernel: *extension* and *elaboration*.

The public announcement of HIV as the cause of AIDS came twenty-three days after the MACS had begun to recruit a cohort. And yet, even as their new study lost its founding purpose, the MACS principal investigators do not recount any sense of crisis. The MACS, in 1984, was in a scientifically enviable position to investigate this new retroviral entity. This is because the MACS kernel had been assembled with the breadth of materials, instruments and experts needed to render HIV a researchable object.

Inspecting each feature of the kernel in turn: the relevant specialists – virologists – were already part of the MACS team; they were readily equipped with the instruments and techniques of their craft; they were already collecting the behavioral data that would become most relevant in understanding transmission in gay men: anal intercourse and intravenous drug use; they were already collecting the material that would most closely come to be associated with transmission: blood.

This is called *repurposing the kernel*: resources and services collected for one purpose are thereafter used for the investigation of new objects of research. For example, blood, collected for the investigation of many possible causes (e.g., it can be screened for pathogens, but can also be used to track drug use or environmental exposures), was repurposed as a key material in the investigation of HIV disease.

The latter is an important point in understanding the capacity of the MACS to adapt to these changes. In the study of cyberinfrastructure, and CSCW studies of scientific work, we have greatly emphasized the importance of data reuse, particularly data in a digital form. In a recent review article, Jirotka, Lee and Olson [18] have called data the "lifeblood of science," a phrase which recurs throughout the literature [32, 40].

Data *are* central to science; the argument in this paper is that the focus on data has thus far come at the expense of a concern with the materiality of scientific resources.

In the study of AIDS, the lifeblood of HIV research has been … blood. In 1985, when the test became available, MACS researchers did not re-inspect their data (nothing to see!), rather they turned to their archives of blood. By retesting these specimens they were able to reclassify the participating men from 'at risk' of contracting AIDS to either HIV-positive or negative – a new category in the columns of their databases. Over time MACS researchers tested the entire specimen archive, producing for each HIV-positive man a new retrospective history of their seroconversion. The blood archive permitted the MACS to push this newly generated diagnostic status back in time [23]: thereafter, for some men, they had always already been infected before the discovery of HIV or the availability of a test.

This is called *elaboration of the kernel:* a change to the cache resulting from the addition of new instruments and categorizations. The sites of collection remained "the same men," but through the HIV test and, essentially, the incorporation of a new variable for each man – serostatus – the study design was transformed in ways that would allow the MACS to continue its investigations following the discovery of HIV. For example, in certain studies the HIV-negative men are used as "controls" to compare with HIV-positive men; they are comparable in that they are demographically similar (e.g., gay, actively sexual, living in the same cities) but with a key difference, serostatus. This comparison enables the isolation of a "natural history of HIV" from the general trajectory of being a gay sexually active American male.

The test is now a routine part of the MACS kernel: it is administered at each six-month visit; the results are carefully catalogued in the data archives. The instrument has been added to the cache, now standardized across the 4 sites.

This point is also crucial in understanding elaboration of the kernel. New instruments cannot simply be "dropped in" to the cache. The HIV test had to be addressed for use in the MACS. In the months and years following the availability of the HIV test, MACS scientists tested the test itself. On the one hand, the validity of the test is important (and scientists report that the early tests produced many false negatives and positives), but for a longitudinal study, the reliability of the test was also critical: ensuring comparability of results (data) over time, and across the 4 sites, is a central service of the MACS. In the end, as with the flow cytometers described above, the MACS chose one HIV test vendor and developed cross-site protocols that sought to routinize the practical execution of the test across sites.

Let us now turn to LTER's refiguration as it began to more centrally include human impact and interrelations with ecology. This is an *extension of the kernel*: the addition of new features to the cache in order to investigate objects of research not originally within the purview of the research infrastructure. A clear example of extension is the addition of urban-ecological sites to LTER.

In 1997, LTER added two new sites of data and specimen collection: the Baltimore Ecosystem Study (BES) and Central Arizona-Phoenix (CAP). With these urban sites came a more systematic integration to LTER of experts focused on, for example, social, psychological and economic objects: social scientists. These new scientists initiated the collection of new data and specimens using new instruments that made urban-ecology researchable, for example: quantifying the coverage of impervious surfaces such as paved roadways; measuring fertilization of lawns with nitrogen-based compounds; or surveying attitudes towards urban parks.

These new sites of collection were different than those of LTER in the past: they were no longer "nature on its own," but instead, as LTER scientists put it, socio-ecological systems. The extensibility of the LTER kernel facilitated the investigation of entirely new objects of research. As LTER scientists have become more interested in the role of humans in ecology, this adaptability of the cache has made new science possible within an old infrastructure. This is technoscientific flexibility.

ADDRESSING: Sustaining and managing change to the kernel

The cache, though, is not "just available," for use. It is worked over continuously to sustain its readiness for future use. As Star reminds us, *infrastructure is relative* [41]: one person's taken for granted resources and services are another person's everyday work of renewing their availability. These are the activities of addressing.

In many senses, the purpose of research infrastructure is to help facilitate technoscientific change, or, as a recent visionary text put it: "revolutionizing science through cyberinfrastructure"[1]. In this paper I have recounted exemplars of such change and how infrastructure supported these new investigations. And yet, these changes are also highly disruptive to scientists and to research infrastructure itself: new theories challenge old; novel instruments are interrogated as producing false findings; and, new methods may leave some scientists feeling like old researchers.

However, in addition to finding tensions within infrastructure there is also long-term cooperation, and this is the key to adapting to new circumstances.

Counter-intuitively, one of the most important strategies for remaining flexible is to keep doing the same thing. More precisely, a central quality of infrastructure is to sustain a consistent level of services and resources. Following the kernel metaphor, these are the activities of *addressing,* in the sense of generating a location within an address space. In the terminology of the kernel, addressing is work to sustain the entanglement of the cache with the kernel.

More concretely, addressing refers to the heterogeneous work, techniques and technologies of those actors that seek to regenerate access to the resources and services of infrastructure. For example, data, particularly data collected somewhere else for some other purpose, are always in danger of losing their meaning and thus their value for reuse in a future investigation. Methods of addressing data may include, for example, careful annotation of who, how and when they were collected [11]**,** or more ambitiously, developing common metadata specifications or semantic representations such as computational ontologies [35]**.**

However, data are only one example: the work of addressing is as diverse and detailed as the range of resources and services that a research infrastructure may avail. Addressing *is* maintenance, repair and upgrade, but these terms are too narrow to cover the full range of activities we find that seek to renew readiness of the cache. We have already seen several other examples of addressing in this paper, e.g., the MACS and LTER regularly organize face-to-face meetings and create online venues, such as forums and webpages, to renew the social ties of their members. We have also seen that instruments (flow cytometers, HIV tests, and water collectors) must be continuously calibrated to ensure comparability. Standardizing is never complete: instruments are forever threatening to "disentangle" and lose their potential to support future research.

The recurrent night terror of disentanglement

In the first ethnography of a scientific laboratory, the observer, now Latour-as-demon, imagined himself trolling through a lab and disentangling samples from the notebooks that gave them meaning:

> Entering the deserted laboratory at night, he opens one of the large refrigerators [.] Each sample on the racks [.] is labeled with a long code number which refers back to the protocol books. Taking each sample in turn, the observer peels off the labels, throws them away and returns the naked samples to the refrigerator. Next morning, he would doubtless witness scenes of extreme confusion. No one would be able to tell which sample was which. It would take up to five, ten, and even fifteen years (the time it took to label the samples) to replace the labels.[24]

Fanciful as this tale may seem, it is a recurrent topic in discussions with the technicians and scientists dedicated to the assembly and care of their archives. In an interview discussing the cold storage of stream water samples, one LTER/BES scientist recounted his own terror:

> That's one long-term study we haven't done: the life of the label glue over time. I dread to think that one day we'll walk into the cold storage room and hundreds of labels will be lying scattered beside the bottles. But I think the extra label tape we put on the lids will hold up.

Labels are the physical point of contact between specimens and the digital databases that track all the details that give those samples meaning. Without a sustained connection to data, the bottles contain mere water, not scientific samples.

Kernel Inversions

Fears of disentanglement do not remain abstractions, they become the topic of what Geoffrey Bowker has called *infrastructural inversions* [5]: a figure /ground reversal in which members turn their attention from their objects of research to the infrastructure that makes that work possible. For instance, in an attempt to strengthen the connection between sample and data, a World Health Organization technical report in 1968 encouraged researchers to use a diamond to directly inscribe identifying codes onto glass vials containing scientific materials rather than relying on glued labels [33]**.**

The link to a label (and thus to data) is only one way that a sample may loose its value. For example, in the MACS, specimens of the blood archive are thawed to check the continued viability of peripheral blood mononuclear cells

(PBMCs – blood cells with a round nucleus, key to immune system function). During a six year-long testing period:

> recently cryopreserved PBMC from HIV-1-infected and HIV-1-uninfected participants at each MACS site were thawed and evaluated. The median recoveries of viable PBMC for HIV-1-infected and -uninfected participants were 80% and 83%, respectively. [2]

Within this study, they found that one of the four MACS sites had a markedly lower recovery rate, and identified the use of "an automated particle counter that was out of calibration and required servicing by the manufacturer." [2]. Such ongoing quality assessment procedures help laboratories improve protocols and performance "to ensure optimal cryopreservation [.] for future studies." [2]**.**

Addressing is an umbrella term for many activities already named within studies of infrastructure, such as maintenance, repair, upgrade, calibration or quality control; but there also many other activities that sustain availability of the cache that, as of yet, have received little scholarly attention.

Protecting and renewing the sites of specimen and data collection is another revealing addressing activity, almost completely unexplored in the literature [though see 21]. From its founding documents, three of the six criteria for funding an LTER site referred to securing the sites of collection: "The principal investigators [.] should consider site integrity, conflict in use of a site, and long-term agreements with site owners." [29]. LTER's sites are (also) real estate that may occasionally change owners and is sometimes transformed into developed land, making impossible further collection of materials.

Even at the urban sites, where human activity is expected, delicate instrumentation can be disturbed. At the streambeds of urban Baltimore, investigators have reported that "kids just love to pee in things!" (interview) such as the metricized plastic tubes they use to measure rainfall. In response, to protect the integrity of the measurement, they have sought to hide these water collectors from the casual passerby.

The MACS reveals a more peculiar form of addressing of its sites of collection, which in biomedical studies is called retention. The participating men can withdraw at any point: they are tied to the MACS only by their willingness to donate their time and bodies to this ongoing enterprise; some men have done so twice a year for nearly thirty years! Keenly aware of the danger of attrition (which could transform the study from a prospective to a retrospective investigation), MACS members work diligently to make it easier for men to participate: i.e., holding the visits in multiple locations in a city to ease the burden of travel; curbing the time and effort men must expend at clinics by truncating questionnaires and minimizing invasive specimen collections; and, by demonstrating their appreciation for the men's volunteered participation by throwing "thank-you" parties at key anniversaries such as the 25th in 2009.

The techniques and technologies described here only begin to scratch the surface of the varieties of addressing work found in these research infrastructures. All of these activities are oriented to preserving availability of resources and services for future investigations. Without them, instruments data, samples, sites, experts and instruments would disentangle; rather than infrastructure they would become suspect measurements, common blood and water, and scattered scientists working on their own.

Can a feature of the cache be meaningfully disentangled from the kernel? Or, more concretely, can data and specimens circulate "outside their infrastructure" without losing their value? This is a complex question, well beyond the scope of this paper, but the short answer is: Yes, but only if they become entangled with another infrastructure. Some data already rely on other infrastructures. Consider the temperature of a stream: such a measurement, taken in degrees centigrade at BES/LTER, can be shared relatively easily across the globe, but this is only because it relies on a an extant global metrological network [30]. Metadata standards [28] can be developed to help data travel, but of course, those data can only do so once they have been redescribed in the metadata specification – a new entanglement with another infrastructure. Further development of this topic will have to await another paper.

SHAGGY DOG AND TRIM POODLE INFRASTRUCTURES

This section compares and contrasts the MACS and LTER kernels, which are revealing of vastly varying possible kernel architectures. The discerning reader may have noticed that this paper has thus far treated the MACS as a whole, while many of the LTER narratives have been drawn from a single site in Baltimore. This is because LTER is a shaggy dog infrastructure and the MACS is a trim poodle.

A "*shaggy dog story"* is one that continues endlessly without coming to a clear point or conclusion[2]. It does not have a punch line. Such stories have an iterative "and then this happened, and then that" quality. Analogously, a shaggy dog infrastructure has fuzzy boundaries, loosely defined membership, and multiple ongoing research approaches across heterogeneous disciplinary fields. A trim poodle infrastructure has much cleaner boundaries: who is or is not a member is clear, what resources are available are commonly defined and its investigations target a circumscribed set of objects.

As Robert Kohler recalls from when he began to study the history of field ecology: "It was interesting but boundless, with few familiar landmarks, and easy to get lost in.

[2] I attribute this colorful language to Geoffrey Bowker, who has long quipped that all infrastructures are shaggy dog stories.

Finding a coherent narrative structure also proved irksome," [21]. This is the case with LTER, which I have studied for many more years than the MACS, but which has proven recalcitrant to a concise narrative. In contrast, the MACS, while still complex and gnarled, immediately lent itself to more succinct storytelling. The histories of these infrastructures recounted above are both vastly figurative, but far more so of LTER than of the MACS.

Shaggy dog and trim poodle are Weberian *ideal types*, they do not accurately describe the real-world complexity of infrastructure. However, ideal types are useful in constructing comparative analysis. We can explain part of these infrastructures' "trim" and "shaggy" qualities by comparing their kernels.

Single-Core and Multi-Core Kernels

LTER has a multi-core kernel, while the MACS has a single-core. A core occurs when efforts are made to centralize, collect or standardize features of the kernel in a common way across time or sites. The concept of a core can help us understand why LTER is a shaggy dog, and the MACS a trim poodle.

Ecology is complex. LTER scientists have innumerable objects of research from soil types to stream chemistry, forest growth and animal populations, from microorganism cultures to watersheds. In contrast, while the MACS certainly investigates many things, generally speaking it has a much narrower set of research objects, and, more importantly, all data and specimens are drawn from the common subject pool of gay and bisexual men in four American cities.

A "core" should not be confused with a site. LTER has twenty-six geographic sites, while the MACS has four. The MACS has a single core because members work to address all their materials, sites of collection and instruments in such as way as to make them comparable. There is an effort to collect the same data, in the same way, at all sites, and to send all of these to a single data management center in Baltimore at Johns Hopkins University. Specimens are transported to a single repository. In turn, all access to data or new requests for specimens go through a single obligatory passage point: the data management center at Johns Hopkins.

Shagginess or trimness is not only a matter of social organization; the objects of investigation will also play a role. Consider, once again, the role of blood in the MACS: it is the material for measuring immune function, for testing the presence of HIV virus itself, for the key markers that (today) transition a person to having AIDS (CD4 counts), and of what has often been called the "golden key" metric of viral load. No such "key" exists for ecology. Shagginess of infrastructure is partially linked to the shagginess of the objects of investigation.

Obviously, efforts to create a single core are never complete. We have seen above how a single instrument out of alignment at one MACS site led to miscounting viable blood cells. There are many other ways that the four sites differ from each other. For example, following the discovery of HIV and an emerging scientific consensus about the centrality of blood for its investigation, the MACS as a whole stopped collecting many specimens, such as stools[3]. However, the Pittsburgh site continued to collect stools for many years afterwards, based on the principal investigator's continuing interest in stomach flora. Generally speaking, in the MACS there is an ongoing effort to generate a single common core to the study; in practice, its realization is a matter of degree. It is this sense that "trim poodle" or "single core" are ideal types.

In contrast, LTER has many cores. The water samples from Baltimore described in this paper, end up in Millbrook, New York. There they are stored with stream samples from that site and with those of the Hubbard Brook site. The methods at Hubbard Brook and Baltimore are ongoingly standardized: by sharing common data and sample protocols, water is collected, analyzed and stored in the same ways for these two sites. In this sense these two sites form a single core for stream water materials. But LTER has many other such cores, e.g., at the Kellogg Biological Station in Michigan they preserve many kinds of plant specimens [6], while at a discontinued LTER site called North Inlet in South Carolina, ecologists still maintain a collection of preserved fish in jars. It is difficult to say with any certainty how many cores LTER may have; a fuzzy kernel is a quality of a shaggy dog infrastructure.

The fuzzy boundaries of the kernel

I once owned a shaggy black dog called Sisco. I enjoyed watching as he would sit on my black shag rug and use his tongue to clean his paws. He often licked a full inch of dark carpet beyond his paw, unaware or indifferent to the limits of Sisco. Shaggy dog infrastructures are this way too.

In an early study of cyberinfrastructure, Charlotte Lee et al. identified the common occurrence of fuzzy membership [25]: a respondent would point her in the direction of another member to interview, but when she arrived, the indicated person would admit no sense of affiliation to the infrastructure. More generally, this is true of all features of the kernel: what is, or is not, part of the cache is easily identified in a trim poodle infrastructure and the source of contestation or confusion in a shaggy dog infrastructure.

[3] Cuts or eliminations to the kernel are called *shedding*, which is as important as the additive activities of elaboration or extension inspected in this article. The kernel does not only grow over time, it is also necessarily – and often controversially – reduced. A further discussion of shedding will appear in forthcoming research.

In terms of membership, it is common for many affiliated researchers, such as graduate students working on LTER projects, to be unaware of any association. In contrast, researchers in the MACS are almost always aware of the use of those subject cohorts and datasets. The case is similar for other features of the cache, e.g., in those LTER sites with very little social scientific investigation (for example, Palmer station in the Antarctic) researchers are less likely to be aware of the methods, instruments and objects of analysis used by social scientists.

Over the years LTER members have sought to address their kernel in order to *trim the shaggy dog*; some of these efforts have been successful, others have vanished without a trace. A most notable success in LTER has been the establishment a network-coordinating center in 1983 to facilitate communication and collaboration across its sites. However, other addressing efforts have come and gone; for example, beginning in 1988 a single Minimum Site Installation (MSI) was implemented in LTER, defining the information technologies that each site should have. This MSI has rarely reappeared in LTER documentation since. A mixed success (thus far) is the effort to develop a common Ecological Metadata Language (EML), which received uneven uptake across the twenty-six sites [28]: today some data are well described in EML, others not at all.

Are the terms shaggy and trim necessarily valuations of these infrastructures? Is trim good and shaggy bad? Certainly, managerial actors tend to prefer a trim structure: standardized data, well catalogued specimens and a clearly defined membership. But shaggy dog infrastructures do 'work.' As Lee et al. note of fuzzy membership: "Participants can successfully accomplish work with a partial view of the organizational membership and structure." There may even be advantages to a shaggy kernel, as with long discussed tension between standardization and flexibility [15]: the loose cross-site coupling found within shaggy dog infrastructures may enable a more nimble adaptation to new research objects or changes to the kernel.

Understanding shaggy and trim qualities, and their implications, requires a much more extensive analysis than is possible in this paper – it is simultaneously a sociotechnical, scientific and institutional matter[4] – but comparing the MACS and LTER kernels has allowed us to begin appreciating their trim and shaggy silhouettes.

DISCUSSION: A kernel approach

The most significant methodological reorientation in a kernel approach is to place supporting the investigation of objects of research at the center of the analysis. The study of infrastructures must be closely tied to the specific research practices they enable, the materials to do so, and the shifting orientation of investigators to novel phenomena. The intellectual contributions of a kernel analysis follow from this reorientation.

In many publications that take cyberinfrastructure as their topic, it is common to find the objects of investigation identified only briefly in the description of the case. Occasionally, objects are altogether absent in such papers and only the disciplines supported are described – as though disciplines can stand in for their objects. Such papers thereafter move on to topics greatly abstracted from the doing of a particular science, e.g., user needs, data sharing, supporting collaboration, disciplinary culture.

A kernel approach insists that the analysis return to actors' research orientation: sharing data about what? collaborating for what purpose? And most importantly, how do these resources and services support investigation (or not)? This paper has made many general points about research infrastructure, but the analysis has always returned to participants' activities of investigating AIDS and ecology.

By returning to the objects of investigation, and tracking their ongoing transformations over time, it has become possible to characterize a particular form of resilience and plasticity thus far largely ignored in studies of research infrastructure: technoscientific flexibility. Infrastructure may be characterized by other forms of flexibility – such as an adaptability to change in its information technologies or its funding arrangements – but a unique feature of both the MACS and LTER is that they have persisted in the face of radical reconfigurations to their scientific methods and objects. More than this: they have continued to conduct investigations within their fields. Without this capacity they would cease to be research infrastructure.

Even while insisting on being specific about objects, the kernel has also served as a new unit of analysis that facilitates a rigorous comparison of infrastructures. We need not ignore the specificity of research objects and practices in order to generalize or to compare. In this paper the leverage of the analysis has come, first, by analyzing infrastructures that offer similar resources and services, and then, through outlining their shaggy and trim qualities, by contrasting the integration of their cores.

This is one possible comparative design. Juxtaposing distinctly different kernel architectures can also be revealing. For example, GEON and LEAD, which I have investigated in the past [36], did not offer their own data and specimens or sites of collection as resources. Instead they gathered the heterogeneous datasets of other enterprises and sought to integrate those data. The service they offered was data interoperation; addressing in such endeavors was rooted in sustaining the links across datasets that scientists used for their inspection of new objects. This is a common architecture within some contemporary

[4] For instance, the MACS has been the beneficiary of many national and international efforts by the institutions of biomedicine to standardize materials and instruments [20].

cyberinfrastructure projects, which emphasize data integration over collection.

Comparing GEON and LEAD's kernel architectures with the MACS and LTER's could be revealing, e.g., on the one hand GEON and LEAD did not need to expend efforts curating data and specimens, nor stewarding the sites of collection, but, on the other hand, data and specimens are a form of scientific capital: without them, such projects may lack the resiliency that comes from holding these repurposable scientific materials.

Thus, a second reorientation in a kernel approach is to be specific about the resources and services that are made available. The kernel is a broad concept; it includes many things traditionally within the purview of infrastructure studies – such as data or maintenance – but it also encourages the analyst to inspect the full range of a cache and its unique forms of addressing. Such an investigation will reveal many surprising resources and ways of sustaining them. Keeping a research infrastructure running is, as we know, annotating data and upgrading a computer, but it is also throwing parties for dedicated volunteers, or hiding instruments from children who want to pee in them.

Other resources that have come to be considered almost synonymous with research infrastructure may altogether fall away in the study of a particular kernel. Consider, for example, the most commonly investigated service of infrastructure (in CSCW): supporting collaboration.

Research infrastructures are often evaluated for their ability to encourage interdisciplinarity or sustain a distributed organization. However, research infrastructures seek to support collaboration only to the extent that its developers consider collaboration to be important in the investigation of new objects. This is not always the case.

For instance, the Open Science Grid (OSG) specifically seeks to minimize the need for human communication. It seeks to support research by making computing power available 'on tap.' Having to go through a human gatekeeper or fill out tedious paperwork in order to gain access to cycles is seen as a hindrance to the doing of science, a waste of time for all members. Instead, the OSG (and grid and cloud computing more generally) is developed with the goal of minimizing the need for scientists to communicate or collaborate: negotiations for access to cycles and storage are largely delegated to the computational system [38]. The hope is that this will free scientists from such tedious activities so they may focus on their research. While coordination is certainly occurring in OSG, we should no more think of this as "collaboration" than you would call pouring water from a tap collaboration with your local waterworks utility.

Finally, the third reorientation in a kernel approach is to recognize that an infrastructure does not simply 'have' resources and services, there are always actors tasked with sustaining, renewing, adding or shedding features of the kernel. Addressing is the work, techniques and technologies that render features *as* resources and services, available for the investigation of new and extant objects.

Notably, while analyzing two thirty-year enterprises, I have not characterized the MACS or LTER as having a 'foundation' that has remained stable across their decades of operation. Neither the cache, nor the objects of investigation have remained consistent across the years. The kernel is more process than thing. I have named only a few of the processes that have changed the kernel in a punctuated manner: elaboration, extension and shedding. There is a great deal more to be written about these changes and certainly many more kinds of such transformations. But the central activity of the kernel is addressing, an ongoing working over of the things of infrastructure (data, specimens, instruments) and people (experts & subjects) to make them the same, comparable, accessible, preserved, or continuing to contribute.

What is not part of the kernel?

At the edges of the kernel and beyond, we will find countless activities, still within the bounds of "an infrastructure," but that have not been part of this analysis. Most notably, one activity that is not "part" of the kernel is an investigation conducted using the resources and services of that infrastructure.

This paper has approached the kernel "from below," focusing on the activities that develop and renew the resources of a research infrastructure. Returning to our guiding metaphor of the kernel of an operating system, what remains to be investigated is approaching the kernel "from above," as a user or "application." To say that the cache is addressed in an effort to regenerate access is not the equivalent of actually gaining access to these resources. How do scientists conducting a study draw on the resources of the kernel, and what challenges do they encounter in practice? For example, in the MACS and LTER, use of the finite materials from the blood and water archive are carefully deliberated. Some resources are available in principle, but in practice are challenging or impossible to access: poorly documented data, or very new data with value for current research that is reserved for core members. Studying how resources are accessed and deployed for particular investigations in practice is one next step in understanding the operations of the kernel.

Conclusion: Evaluating research infrastructure

This paper has presented a framework for understanding how old organizations have sought to renew themselves to sustain value for new science. This comparative investigation of the past and the present of two research infrastructures has sought to understand what kinds of changes they have encountered over time, and what strategies they have employed in the face of transformations to the landscape of their science. MACS and LTER are active today, continuing to investigate HIV disease and global ecologies; they also continue to plan and organize for a future that promises many ongoing alterations to their research landscapes.

We can expect but not precisely predict continuing transformations to the scientific ecologies of today's research infrastructure. New research infrastructure projects, i.e., cyberinfrastructure, will face a different and unique set of challenges than those of the past thirty years. From our vantage point, we can only sketch these oncoming scientific changes, but by investigating the past and present of infrastructures that have weathered transformations, this research has sought to inform future infrastructure development efforts by articulating strategies of the long-term: organizational forms and methods of design with a track-record of facilitating responsiveness to change.

Considered in this manner – framing research infrastructure as the repurposing, elaborating and extending the kernel, could serve to develop methods for evaluating the ability of research infrastructure to support future research. Defining 'progress in science' is a notoriously fraught philosophical endeavor, but evaluating a research infrastructure according to its ability to support the investigation of new objects could serve as a new metric: What does it take to gain access to data and specimens for repurposing in new investigations? How much work does it take to add a new instrument? Are seasoned scientists and diverse specialists easily available to help navigate the use of resources and services?

Posing the evaluation of research infrastructure in this manner leads to a deeper, more troubling question: are there objects of research that a kernel architecture makes wholly impossible to investigate? Consider: the MACS is rooted in the study of American bodies. While there are certainly many similarities between HIV disease in the U.S. and the natural and treated histories of HIV around the globe, today we also know that AIDS manifests in distinctly different ways in the developing world. The MACS has contributed enormously to our knowledge of HIV disease, and to our ability to manage that illness; but our understanding of specific variation in the developing world has come from other investigations. By virtue of their architecture, there are limits to the flexibility of any kernel that may render entire domains of inquiry opaque.

ACKNOWLEDGEMENTS

Jessica Beth Polk was an invaluable collaborator and field worker during investigations of the MACS, and in the development of the kernel concept. This paper was presented at a workshop in the Department of Informatics at the University of Oslo; I would like to thank the participants at that event for their feedback and to Jennifer Feldman for her insightful comments on final revisions. This research has been supported by NSF/VOSS award #0904145 'Delegating Organizational Work to Virtual Organization Technologies.'

REFERENCES

1. Atkins, D. E. C. *Revolutionizing Science and Engineering Through Cyberinfrastructure*. National Science Foundation, Washington, DC, (2003).
2. Aziz, N., Margolick, J. B., Detels, R., Rinaldo, C. R., Phair, J., Jamieson, B. D. and Butch, A. W. Value of a quality assessment program in optimizing cryopreservation of peripheral blood mononuclear cells in a multicenter study. *Clinical and Vaccine Immunology,* 20, 4 (2013), 590-5.
3. Baker, K. S. and Yarmey, L. Data stewardship: Environmental data curation and web-of-repositories. *International Journal of Digital Curation*, 4, 2 (2009), 1-12.
4. Birnholtz, J. P. and Bietz, M. J. Data at work: supporting sharing in science and engineering. *GROUP (2003), 339-348*.
5. Bowker, G. C. *Science on the Run: Information Management and Industrial Geophysics at Schlumberger, 1920-1940*. MIT 1994.
6. Burton, M. and Jackson, S. J. *Constancy and Change in Scientific Collaboration: Coherence and Integrity in Long-Term Ecological Data Production*. HICSS, (2012), 353-362.
7. Charmaz, K. *Constructing grounded theory*. Sage (2006).
8. Dourish, P. and Edwards, W. K. A Tale of Two Toolkits: Relating Infrastructure and Use in Flexible CSCW Toolkits. *JCSCW* 9, 1 (2000), 33-51.
9. Edwards, P. N. *A vast machine: computer models, climate data, and the politics of global warming*. MIT, 2010.
10. Edwards, P. N., Jackson, S. J., Chalmers, M. K., Bowker, G. C., Borgman, C. L., Ribes, D., Burton, M. and Calvert, S. *Knowledge Infrastructures: Intellectual Frameworks and Research Challenges*. Deep Blue, Ann Arbor, 2013.
11. Edwards, P. N., Mayernik, M. S., Batcheller, A. L., Bowker, G. C. and Borgman, C. L. Science friction: Data, metadata, and collaboration. *Social Studies of Science*, 41, 5 (2011), 667-690.
12. Faniel, I. M. and Jacobsen, T. E. Reusing Scientific Data: How Earthquake Engineering Researcher Assses the Reusabilty of Colleagues' Data. *JCSCW*, 19, 3-4 (2010), 355-375.
13. Finholt, T. A. Collaboratories. *Annual Review of Information Science and Technology*, 36, (2004), 73-107.
14. Glaser, B. G. and Strauss, A. *The discovery of grounded theory: strategies for qualitative research*. Aldine, 1973.
15. Hanseth, O., Monteiro, E. and Hatling, M. Developing Information Infrastructure: The Tension between Standardization and Flexibility. *Science, Technology & Human Values*, 21, 4 (1996), 407-426.

16. Howison, J. and Herbsleb, J. D. *Incentives and integration in scientific software production*. CSCW (2013), 459-470.
17. Jackson, S. J. and Barbrow, S. Infrastructure and vocation: field, calling and computation in ecology. *CHI* (2013), 2873-2882.
18. Jirotka, M., Lee, C. P. and Olson, G. M. Supporting Scientific Collaboration: Methods, Tools and Concepts. *JCSCW* (2013 online first), 1-49.
19. Keating, P. and Cambrosio, A. *Interlaboratory life: regulating flow cytometry*. The Invisible Industrialist: Manufacturers and the Construction of Scientific Knowledge. MacMillan (1998) 250-95.
20. Keating, P. and Cambrosio, A. *Biomedical platforms*. MIT Press, (2003).
21. Kohler, R. E. *Landscapes and labscapes: Exploring the lab-field border in biology*. University of Chicago Press, (2002).
22. Latour, B. *Circulating reference: Sampling the soil in the Amazon forest*. Pandora's Hope. Harvard Press, Cambridge, (1999), 24-79.
23. Latour, B. On the partial existence of existing and nonexisting objects. *Biographies of scientific objects,* University of Chicago Press (2000), 247-269.
24. Latour, B. and Woolgar, S. *Laboratory life: the construction of scientific facts*. Sage Publications, Beverly Hills, 1979.
25. Lee, C. P., Dourish, P. and Mark, G. *The human infrastructure of cyberinfrastructure*. CSCW (2006) 483-92.
26. Lee, E. S., McDonald, D. W., Anderson, N. and Tarczy-Hornoch, P. Incorporating collaboratory concepts into informatics in support of translational interdisciplinary biomedical research. *International journal of medical informatics*, 78, 1 (2009), 10-21.
27. LTER *A long-range Strategic Plan for the Long Term Ecological Research Network*. LTER Network Office, 1989.
28. Millerand, F., Ribes, D., Baker, K. S. and Bowker, G. C. Making an Issue out of a Standard: Storytelling Practices in a Scientific Community *Science, Technology & Human Values*, 38, 1 (2013), 7-43.
29. NSF. *LTER Request for Proposals*. 1980.
30. O'Connel, J. Metrology: 'The Creation of Universality by the Circulation of Particulars'. *Social Studies of Science*, 23 (1993), 129-173.
31. Olson, G. M., Zimmerman, A. and Bos, N. *Scientific collaboration on the internet*. MIT Press, 2008.
32. Onsrud, H. J. and Campbell, J. Big opportunities in access to "Small Science" data. *Data Science Journal*, 6 (2007), 58-66.
33. Radin, J. Latent Life: Concepts and Practices of Human Tissue Preservation in the International Biological Program. *Social Studies of Science* (Online First).
34. Ranon, R., Marco, L. D., Senerchia, A., Gabrielli, S., Chittaro, L., Pugliese, R., Cano, L. D., Asnicar, F. and Prica, M. *A Web-based Tool for Collaborative Access to Scientific Instruments in Cyberinfrastructures*. Grid Enabled Remote Instrumentation. Springer US, (2009).
35. Ribes, D. and Bowker, G. C. Between meaning and machine: learning to represent the knowledge of communities. *Information and Organization*, 19, 4 (2009), 199-217.
36. Ribes, D. and Finholt, T. A. The long now of technology infrastructure: Articulating tensions in development. *Journal for the Association of Information Systems (JAIS): Special issue on eInfrastructures*, 10, 5 (2009), 375-398.
37. Ribes, D. and Jackson, S. J. *Data Bite Man: The Work of Sustaining a Long-Term Study*. "Raw Data" is an Oxymoron. MIT Press (2013) 147-166.
38. Ribes, D., Jackson, S. J., Geiger, R. S., Burton, M. and Finholt, T. Artifacts that organize: Delegation in the distributed organization. *Information and Organization*, 23, 1 (2012), 1-14.
39. Ribes, D. and Lee, C. P. Sociotechnical Studies of Cyberinfrastructure and e-Research: Current Themes and Future Trajectories. *JCSCW* 19, 3-4 (2010), 231-244.
40. Ribes, D. and Polk, J. B. Historical Ontology and Infrastructure. *iConference* (2012), 252-264.
41. Star, S. L. and Ruhleder, K. Steps Toward an Ecology of Infrastructure: Design and Access for Large Information Systems. *Information Systems Research*, 7, 1 (1996), 111-134.
42. Vertesi, J. and Dourish, P. *The value of data: considering the context of production in data economies*. CSCW, 2011, 533-542.