The Biggest Data of All:
Making and Sharing a Digital Universe

*by W. Patrick McCray**

ABSTRACT

Throughout the twentieth century, astronomers moved from and between different data-collection regimes, from the photographic to the electronic and, finally, to the born-digital era. At the same time, the focus of scientific discovery shifted away from the telescope itself to the hard drive, the database, and digital archives. This essay builds on the assumption that the sharing and circulation of astronomical data—as with other kinds of scientific data—have become core research activities that demand an increasing fraction of researchers' time, money, and expertise. The examples presented here give insights into the larger and gradual digitization process that unfolded throughout the entire international astronomy community. Although the examples chosen here depict local processes, the importance of sharing digital data transcended specific institutions, individual research questions, and national boundaries.

> As if an astronomical observatory should be made without any windows and the astronomer within should arrange the starry universe solely by pen, ink, and paper.
>
> —Charles Dickens, *Hard Times*, 1854[1]

To neophytes, doing astronomical research might appear relatively straightforward. When a science writer asked Wallace Sargent what he would do with better and bigger instruments, he responded that he would "point the fucking telescope at the sky and see what's out there."[2] If taken at face value, the noted observational astronomer's pro-

* Department of History, University of California, Santa Barbara, CA 93106-9410; pmccray@history.ucsb.edu.

Thanks to the Huntington Library and the California Institute of Technology for providing opportunities to do the initial research for this essay. The Max Planck Institute for the History of Science contributed a short-term residency in 2014 that enabled ideas here to coalesce further. David Brock, David DeVorkin, Alison Doane, Josh Grindlay, Richard S. Ellis, George Djorgovski, Robert Hanisch, Robert Kirshner, and Donald C. Wells generously offered thoughts and recollections. Finally, I wish to thank the other contributors to this volume and *Osiris*'s anonymous reviewers for their helpful suggestions.

[1] Dickens, *Hard Times* (London, 1854), 289.

[2] Ann Finkbeiner, *A Grand and Bold Thing: An Extraordinary New Map of the Universe Ushering in a New Era of Discovery* (New York, 2010), 1. I interviewed Wallace Sargent twice myself, once in February 2003 (https://www.aip.org/history-programs/niels-bohr-library/oral-histories/31826-1 [accessed 15 September 2014]) and again in February 2012, about eight months before his death on 29 October 2012. Based on our conversations, his colorful statement reflected his personal vision of observational astronomical practice.

fane reply implies a traditional mode of research anchored to the telescope. At about the same time, another (younger) researcher presented a different picture. “I do not believe,” he said, “discoveries are made at the telescope riding the weather variations like a cowboy riding a bucking bronco.” Instead, they emerge “after long nights at a terminal trying to reconcile an awkward data set with preconceived models.”[3]

Over the last four decades, observational astronomers’ locus for discovery has expanded to include sites other than the telescope; these include the hard drive, the database, and the digital data archive. This essay builds on the assumption that sharing and circulating astronomical data—as opposed to collecting it—have become core research activities that demand an increasing fraction of researchers’ time, money, and expertise. Astronomers, like their counterparts in biology, meteorology, and other fields, have increasingly been obliged to become data scientists as well.

In the past half century, astronomers encountered different and often overlapping data regimes. In the photographic era—its origins in the mid-nineteenth century were roughly contemporaneous with those of astrophysics—raw data was collected via photographic means, and it remained photographic. Adjoining and overlapping this, starting shortly before World War II, was an electronic era. Here, devices such as photomultipliers and image tubes augmented and complemented established photographic techniques. However, data was still recorded in analog fashion on strip charts or punch cards that a person would later analyze. Note that “computerization” is not necessarily equivalent to “digitization”; that is, data produced via electronic instruments did not have to terminate in a digital format.[4] Eventually, astronomical practice gradually moved into a born-digital era where the raw data—itself a contested term—was collected, manipulated, and stored in a digital format.[5] The boundaries between these three eras were blurred and indistinct as older technologies and techniques endured and complemented newer modes.

And what of this digital astronomical data? Some experts have claimed that astronomical data differs from other kinds of scientific or commercial data. James N. Gray, a computer software scientist and data management expert, often joked with colleagues that he and other data experts “liked working with astronomers because their data is worthless.”[6] By this, Gray meant that astronomical data has little commercial value (a point that can be compared with Dan Bouk’s comments on the value of data in this volume).[7] And, unlike the massive databases maintained by corporations, health enterprises, or government agencies, astronomers’ data poses no legal or ethical implications or privacy constraints. And although astronomers’ data has little relevance

[3] Matt Mountain, “New Observing Modes for the 21st Century: A Summary,” in *New Observing Modes for the Next Century*, ed. Todd Boroson, John Davies, and Ian Robson (San Francisco, 1996), 235–44, on 240.

[4] See Nathan Ensmenger, “The Digital Construction of Technology: Rethinking the History of Computers in Society,” *Tech. & Cult.* 53 (2012): 753–76, in which Ensmenger highlights the distinction between computerization and digitization.

[5] Lisa Gitelman, ed., “*Raw Data*” *Is an Oxymoron* (Cambridge, 2013).

[6] Gray, whose 2007 disappearance at sea prompted a massive high-tech search, often deployed this anecdote in his public presentations. It appears in a number of places, including a tribute to Gray from a colleague: Alexander S. Szalay, “Jim Gray, Astronomer,” *Comm. ACM* 51 (2008): 58–65. Gray’s talks and papers are available at http://research.microsoft.com/en-us/um/people/gray/ (accessed 15 September 2014).

[7] Dan Bouk, “The History and Political Economy of Personal Data over the Last Two Centuries in Three Acts,” in this volume.

for national security (unlike census or climate data), it correspondingly has few political or policy implications.[8]

Besides possessing the ability to be shared with relative ease, astronomical data is derived from actual observations. It is, in other words, both "real" and "well-documented, spatially and temporally."[9] As data goes, however, it isn't perfect. Cosmic rays, airplanes passing overhead, and poor weather can make astronomical data "dirty," which, to some computer experts, presents intriguing challenges. Moreover, the questions one can ask of the data are scientifically interesting. These differences in degree, if not kind, are something to consider when comparing astronomical data to other kinds of collected data—genomic, biological, environmental—as well as data generated in a lab, differences that other authors in this volume consider.

Astronomers work in a profoundly data-rich world. For example, since its 1990 launch, the Hubble Space Telescope (HST) alone has transmitted dozens of terabytes of observational data back to earth. The overabundance of data, in fact, presented scientific communities with tremendous challenges.[10] But, during the 1970s, astronomers began noting an especially significant discontinuity. It is here that one begins to find astronomers' desperate-sounding references to floods, deluges, and explosions of data. A British committee reported, for example, that "data generated by powerful new detectors . . . are overwhelming," while American astronomers grudgingly accepted the "potential disruption" of computers in the "quiet austerity of a telescope dome," because scientists simply "cannot cope with the large amount of raw data produced by electronic detection systems."[11] Where once astronomers complained that they lacked sufficient data, they now started to worry about drowning in it.

There were, however, considerable obstacles associated with converting, sharing, processing, and archiving scientific data. Paul Edwards uses the metaphor of "data friction"—a term adopted in this volume by contributors like David Sepkoski and Etienne Benson—to explain the "costs in time, energy, and attention" needed to "collect, check, store, move, receive, and access data."[12]

[8] Gray's comment only applies to certain types of contemporary astronomical data. In the past, the military services were keenly interested in particular types of astronomical information. This, of course, connects to broader discussions of the nature of scientific objectivity itself; Lorraine Daston and Peter Galison, *Objectivity* (New York, 2007).

[9] Alexander S. Szalay, "Publishing Large Datasets in Astronomy—the Virtual Observatory," in *Electronic Scientific, Technical, and Medical Journal Publishing and Its Implications: Proceedings of a Symposium*, ed. Technical Committee on Electronic Scientific, and Medical Journal Publishing (Washington, D.C., 2004), 83–6, on 85.

[10] This sense of crisis was by no means a new phenomenon; see Daniel Rosenberg, "Early Modern Information Overload," *J. Hist. Ideas* 64 (2003): 1–9; as well as Daniel R. Headrick, *When Information Came of Age: Technologies of Knowledge in the Age of Reason and Revolution, 1700–1850* (Oxford, 2000); Lars Heide, *Punched-Card Systems and the Early Information Explosion, 1880–1945* (Baltimore, 2009).

[11] "Report of the Panel on Astronomical Image and Data Processing," March 1979, personal papers of Richard S. Ellis (hereafter cited as "RSE"); copies in W. Patrick McCray's possession. Lloyd B. Robinson, "On-Line Computers for Telescope Control and Data Handling," *Annu. Rev. Astron. Astrophys.* 13 (1975): 165–85, on 175. See also Stephen E. Strom to Peter Boyce, 6 July 1977, personal papers of Donald C. Wells (hereafter cited as "DCW"); copies in McCray's possession.

[12] Paul N. Edwards, *A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming* (Cambridge, 2010), 84. See also Edwards, Matthew S. Mayernik, Archer L. Batcheller, Geoffrey C. Bowker, and Christine L. Borgman, "Science Friction: Data, Metadata, and Collaboration," *Soc. Stud. Sci.* 41 (2011): 667–90; David Sepkoski, "The Database before the Computer?"; and Etienne Benson, "A Centrifuge of Calculation: Managing Data and Enthusiasm in Early Twentieth-Century Bird Banding," both in this volume.

In addition to data friction, we also encounter social friction. Whether it was analog or digital, collecting, analyzing, and sharing astronomical data required that considerable work be performed. Different communities of technical and scientific experts were implicated in the transnational project of constructing a digital facsimile of the universe. Goals and methods were not always aligned, and not all researchers were willing to readily cede primacy to the data archive instead of the telescope. At its core, this essay examines ways in which these many and varied points of friction were successfully greased or remained stubbornly sticky.

This essay extends our understanding of modern astronomical practice beyond the telescope itself. Much of the history of astronomy, at least where it intersects with technology, has focused on the building of institutions and instruments along with the politics and patronage that made this possible.[13] Less attention has been paid to astronomy's "knowledge infrastructure" in which the production and circulation of images is central.[14] But the data—once analog in form but now almost always digital—is indispensable for producing new knowledge about the universe. This essay, in other words, seeks to better understand what happens at the other end of the telescope.

Finally, the practices and activities associated with the world of astronomical data offer an opportunity to think more directly about the economies associated with astrophysics and modern science in general. There are, of course, issues of political economy as scientists maneuver to secure the resources necessary for building increasingly costly and complex instruments. But there is also the moral economy of astronomy to consider.[15] When interviewing astronomers over the age of fifty or so, it is not hard to elicit stories about colleagues with offices full of exposed photographic plates or data tapes that remained unanalyzed and unpublished—what one scientist referred to as the "mine, mine, mine syndrome."[16] While perhaps apocryphal, such tales are instructive in that they are often presented as a critique of scientists' behaviors. These two differ-

[13] See, e.g., Richard Hirsh, *Glimpsing an Invisible Universe: The Emergence of X-Ray Astronomy* (Cambridge, 1983); David H. DeVorkin, *Science with a Vengeance: How the Military Created the US Space Sciences after World War II* (New York, 1992); W. Patrick McCray, *Giant Telescopes: Astronomical Ambition and the Promise of Technology* (Cambridge, Mass., 2004).

[14] Astronomers, perhaps reflecting their long visual tradition, routinely refer to their data as images. While this term might in some cases refer to familiar pictures of stars and galaxies, more often these data are astronomical spectra that yield information about an object's composition, temperature, and other physical conditions. The term "knowledge infrastructure" is adapted from Christine L. Borgman's *Big Data, Little Data, No Data: Scholarship in the Networked World* (Cambridge, 2015), 4.

[15] In my usage, a moral economy refers to the values and beliefs associated with the production of knowledge and the circulation of resources. These values are not fixed but instead change over time and vary from place to place. New technologies (or changes to existing ones) can also modify the moral economy in which scientists and engineers do their work. The classic articulation of the concept of a moral economy remains E. P. (Edward Palmer) Thompson's "The Moral Economy of the English Crowd in the Eighteenth Century," *Past and Present*, 50 (1971): 76–136, as well as his earlier book, *The Making of the English Working Class* (Vintage, 1963). Historians of science have deployed the idea in several different ways. See, e.g., Steven Shapin, "The House of Experiment in Seventeenth-Century England," *Isis* 79 (1988): 373–404; Robert E. Kohler, *Lords of the Fly:* Drosophila *Genetics and the Experimental Life* (Chicago, 1994); Lorraine Daston, "The Moral Economy of Science," *Osiris* 10 (1995): 2–24; W. Patrick McCray, "Large Telescopes and the Moral Economy of Recent Astronomy," *Soc. Stud. Sci.* 30 (2000): 685–711. More recently, moral economy has been used as a framework for exploring the behavior of scientific communities in terms of credit and authorship; see, e.g., Bruno J. Strasser, "The Experimenter's Museum: GenBank, Natural History, and the Moral Economies of Biomedicine," *Isis* 102 (2011): 60–96.

[16] Finkbeiner, *A Grand and Bold Thing* (cit. n. 2), 44.

ent economies are not separate spheres but rather overlapping regimes in which the circulation and ownership of resources—money, data, credit—are central.

Compared with relatively new subfields like radio or X-ray astronomy—fields that emerged after World War II and carried less historical baggage—the traditional optical astronomy community had more deeply ingrained data practices. Astronomers observing in optical wavelengths had used photographic methods to collect data and record images and spectra since the mid-nineteenth century. But in a relatively short span of time, between roughly 1960 and 1980, observational astronomers' view of the sky transformed to a digital one. Not surprisingly, changes in the moral economy were most contested and the "coefficient of friction" highest in this community. Accordingly, this is where I have focused most of my attention.[17]

This essay uses several examples as probes to explore the social life of astronomical data. Instead of focusing on one specific institution or national context, I have opted instead to present illustrative snapshots that represent a diverse community of actors and institutions. However, a common meta-theme connects these activities: circulation. One could argue that astronomers—even if they did not always articulate it directly as such—directed their efforts toward the broader goal of reducing friction so that data could move about and between researchers.

The digitization of astronomy and its effects on practices like data sharing was not restricted to one country or disciplinary subfield. Rather, it was a process that instrument builders, observers, and theoreticians alike experienced in some way. Nonetheless, the examples presented here give insights into larger data-centric processes that unfolded throughout the entire international astronomy community. This approach allows us to see the transformation at several different scales, from the local context of individual laboratories to the transnational circulation of data. Although the examples chosen here depict local processes, the importance of sharing digital data transcended specific institutions, individual research questions, and national boundaries. For all astronomers, it was, in both senses of the phrase, a universal concern.

CONVERSION EXPERIENCES

In August 1970, astronomers and engineers convened at the Royal Observatory in Edinburgh to discuss best practices for applying automation techniques to astronomical data. The Edinburgh meeting occurred at a time when there was a good deal of introspection concerning the relationship between scientists and the data they collected. Astronomers' anxiety and enthusiasm mirrored feelings expressed earlier by high-energy physicists when particle accelerators and bubble chambers grew enormously in complexity and cost as well as the amount of data generated.

Despite one scientist's warning—if "we rely on automats to do everything . . . there will be no more men in the full sense"—at the Edinburgh meeting the mood was gen-

[17] Data collected at optical wavelengths with ground-based or space-based telescopes, alone or combined with data from other wavelength regimes such as radio or X-ray, still results in the majority of research publications; Helmut Abt, "The Most Frequently Cited Astronomical Papers Published during the Past Decade," *Bull. Amer. Astron. Soc.* 32 (2000): 937–41. Since 1970, astronomers have become less parochial in terms of the wavelength regimes (optical, infrared, radio, etc.) in which they worked; Abt, "The Growth of Multi-Wavelength Astrophysics," *Publ. Astron. Soc. Pacif.* 105 (1993): 437–9.

erally optimistic.[18] According to Jaap Tinbergen, a Dutch scientist who had recently migrated from radio astronomy to optical wavelengths, one device appeared as "an extremely adult sort of marriage" between computer and mechanical techniques.[19] Tinbergen directed his compliment toward an ambitious prototype machine, built and promoted by Edward J. Kibblewhite, which could automatically scan and measure astronomical photographs.

In 1970, Kibblewhite was a twenty-six-year-old graduate student just a few months away from filing his dissertation at Cambridge University. Like many in his professional cohort, Kibblewhite had moved into astronomy from another field (in his case, it was electrical sciences). Driving this demographic shift was the growing sophistication of electronic instrumentation used to collect data. As in other fields, research using optical telescopes—as was already the case for emerging fields like radio astronomy—became increasingly dependent on instrumentation developed by "gadgeteers" and "electronickers."[20]

In 1966, Kibblewhite proposed building an "automatic Schmidt reduction engine" for his dissertation.[21] Its prime application would be analyzing images taken with Schmidt telescopes, instruments whose optics are designed to take in much wider fields of view compared with conventional telescopes. To understand the data challenge posed by these large-scale survey telescopes, consider their output. The photograph itself is a negative—bright objects like nearby stars appear as dark black spots while galaxies are fainter and fuzzier. A typical exposure might contain as many as one million astronomical objects. About half of these might be stars, the other half galaxies. In addition, the chemical emulsion that records the images has an inherent graininess. Thus, distinguishing between very faint stars, distant galaxies, and the emulsion background itself posed a challenge. Another factor was the sheer amount of data each photographic plate contained. Some two billion "individual picture points" required sifting in order to find "data of interest."[22]

For the next five years, Kibblewhite designed and built what eventually became the Automated Photographic Measuring facility (or APM).[23] The initial cost was just under £33,000, a considerable sum in the late 1960s (roughly US$800,000 in 2016).[24] In developing his design, Kibblewhite looked to previous machines as something to improve upon. For example, astronomers had routinely used "measuring engines" since the 1950s. These instruments scanned photographs and electronically recorded information such as the coordinates of stars and galaxies.[25] Commercial firms like Perkin-

[18] Quote from Jean Rösch, "Introductory Address," in *Automation in Optical Astrophysics*, ed. H. Seddon and M. J. Smyth (Edinburgh, 1971), 3–7, on 7.

[19] Discussion following Edward J. Kibblewhite, "The Cambridge Automatic Plate-Measuring Project," in Seddon and Smyth, *Automation* (cit. n. 18), 122–3, on 123.

[20] Leo Goldberg to Jesse L. Greenstein, 15 October 1958, Box 12, "Goldberg" folder, and Greenstein to Alec Boksenberg, 17 September 1975, both from Jesse L. Greenstein papers, Archives of the California Institute of Technology, Pasadena.

[21] Letter and memo from Kibblewhite to Roderick Redman, 20 June 1966, personal papers of Edward J. Kibblewhite (hereafter cited as "EJK"); copies in W. Patrick McCray's possession.

[22] These data challenges are described in Ed Kibblewhite, "Counting the Stars by Computer," *New Scientist* 99 (1983): 478–82.

[23] There are various names ascribed to the acronym in the literature. In addition to Automatic Photographic Measuring facility, there is the Advanced Plate Measuring facility.

[24] Edward Kibblewhite, "The Design of Automatic Systems for the Analysis of Astronomical Photographic Data," unpublished report, 11 July 1967, EJK.

[25] John Lentz and Richard Bennett, "Automatic Measurement of Star Positions," *Electronics* 27 (1954): 158–63; S. Vasilevskis, "Automatic Measurement of Astrographic Plates," *Astron. J.* 62 (1957): 208–12.

Elmer eventually made "microphotometers" that allowed researchers to manually map the location of a star or galaxy on a photographic plate, measure its optical density, and convert the signal into a value representing the object's brightness.[26]

In 1970, the most sophisticated scanning machine for astronomy to date was located at the Royal Observatory, Edinburgh. The GALAXY machine—a tortured acronym for "General Automatic Luminosity and X-Y" measuring engine—was first proposed in the late 1950s by Peter B. Fellgett, a professor of Cybernetics and Instrument Physics at the University of Reading.[27] Development and testing took more than a decade, and the finished machine did not start operation until 1969. Fellgett's idea was to shine a light beam through a photographic plate. The transmitted signal, recorded on the other side, was diminished when the light passed through an image of a star or galaxy. The automated machine would then register the object and record its position and magnitude.

However, the GALAXY machine projected a relatively faint spot from a cathode ray tube and could only record about 1,000 objects per hour from a photographic plate. In contrast, Kibblewhite decided to use a bright laser beam as the light source for his APM. When rapidly moved, the scanner could process an entire Schmidt plate about a hundred times faster and do so automatically.[28]

Besides the optomechanical parts of the scanner, Kibblewhite and the small team working on the APM built a system to handle the large amount of data their machine generated. For the actual image analysis, Kibblewhite found assistance from an unexpected source. In 1967, he met a cancer researcher at Cambridge's Pathology Department who was developing software to process images of cell nuclei. When stained black, biological cells, Kibblewhite wrote, "looked just like star images."[29] The ability to subtract the background as well as delineate the edges of "fuzzy" objects was essential for researchers in both biology and astronomy, and Kibblewhite adopted a variation of this biology software for his APM.

The APM was interactive in the sense that a user could monitor the data conversion process in real time. It could also combine data from a number of different photographic plates. This was important because astronomers often observe the same patch of sky using different filters. As a result, only certain wavelengths were recorded on each photographic plate. Comparing plates made with different "colors" allowed one to distinguish, for example, between stars and other objects such as quasars.

This feature enabled Kibblewhite and a small group of scientists in 1987 to publish the discovery of the first quasar with a redshift of 4, making it the most distant object seen in the universe at the time.[30] How the APM was used by astronomers is also telling. The 1987 quasar paper, for example, used data—in this case, photographs from the UK's Schmidt Telescope facility in Australia—collected by another observer. The actual discovery, in other words, occurred after the data was reanalyzed using the APM, a point that reinforced the potential of working with archived data.

[26] For a good overview of these machines' history as well as technical details, see http://www.astro.virginia.edu/~rjp0i/museum/index.html (accessed 30 June 2014).

[27] G. S. Walker, "The Design and Development of the GALAXY Machine," in Seddon and Smyth, *Automation* (cit. n. 18), 103–8.

[28] The entire system is described in Edward J. Kibblewhite, "The Automatic Measurement of Astronomical Photographs" (PhD thesis, Univ. of Cambridge, 1971).

[29] Kibblewhite, "Counting the Stars" (cit. n. 22).

[30] S. J. Warren, P. C. Hewett, M. J. Irwin, R. G. McMahon, M. T. Bridgeland, P. S. Bunclark, and E. J. Kibblewhite, "First Observation of a Quasar with a Redshift of 4," *Nature* 325 (1987): 131–3.

Kibblewhite described the APM as a "national facility" available to "astronomers from all over the world" who wanted to convert and analyze their photographic data.[31] Scientists could come to Cambridge with their own astronomical photographs, and once the conversion was done—it took about seven hours to convert a typical plate into about two billion digital pixels—the astronomer could then "walk away with his data and start working out what it all means."[32] If two people wanted to study the same astronomical objects, "we would try to persuade them to collaborate," Kibblewhite recalled. But, if they were reluctant, "we would scan the same plates with the same selection criteria and provide them each with their own, but different, magnetic tape."[33] Established rules of sharing and ownership prevailed as converted data still belonged to the individual scientist rather than going to a common repository for later use by another person. To be fair, a major reason was technical, not cultural. There simply was not enough computer memory available for the APM project to retain copies, so the data "archive" still resided in the original photographic plates.

By the end of the 1970s, astronomical data could be rendered digital, either via conversion or at the moment of its creation via a veritable zoo of increasingly sophisticated instruments. These technological developments compelled astronomers, electrical engineers, and software writers to collaborate with one another more often.[34] Besides fostering increased need for collaboration and an expanded professional skill set, the digital nature of astronomical data raised an increasingly important issue. As opposed to the physical artifacts that characterized the photographic era, once data was digital, it became—at least in principle—more portable. As Joanna Radin's essay in this volume notes, one attribute of "Big Data" is its ability to "radically transcend" the "locality of its production."[35] Scientific data that could circulate more easily had the potential to more thoroughly disrupt long-standing community traditions and norms about ownership and access. But in order for astronomers to circulate their data more easily, they had to be able to share it.

"*ANY* WELL DEFINED FORMAT IS INFINITELY PREFERABLE TO NONE"

Imagine it is 1976 and you are an observational astronomer. Regardless of what kind of telescope you used—optical or radio, public or private, orbiting in space or sitting on a mountaintop—if you wanted to share your data, it was hard to do. In the older analog tradition, astronomers might loan photographic plates to colleagues, and observatories maintained physical libraries of the same. But, as more data was born-digital or converted to a digital format, the ability to share it posed an increasingly problematic issue.

Several factors contributed to astronomers' growing concern about their data. The prime driver was the "swelling flood of data" that astronomers' nightly observing runs produced.[36] Scientists also lacked appropriate tools to tackle the "daunting task" of

[31] Kibblewhite, "Counting the Stars" (cit. n. 22).
[32] Ros Herman, "Starmap Calibration Goes Automatic," *New Scientist* 84 (1979): 522–3.
[33] Edward J. Kibblewhite, personal correspondence with W. Patrick McCray, 25 June 2014.
[34] W. Patrick McCray, "How the Astronomers' Sky Became Digital," *Tech. & Cult.* 55 (2014): 908–44.
[35] Joanna Radin, "'Digital Natives': How Medical and Indigenous Histories Matter for Big Data," in this volume.
[36] M. J. Disney and P. T. Wallace, "STARLINK," *Quart. J. Roy. Astron. Soc.* 23 (1982): 485–504.

turning data into an "astrophysically useful form."[37] So, despite the growing capabilities of new digital detectors and instruments, astronomers lagged in their ability to "extract and study the relevant bits from this mass of data."[38] Greater challenges lay ahead. In the late 1970s, scientists in the United States, the United Kingdom, and Europe awaited the launch of what became the HST. The "immense amount of data" generated by Hubble and other space-based facilities meant the existing data glut would be "greatly aggravated" while a "lot of valuable science could be lost" because of poor data-handling capabilities.[39]

At the same time, astronomers' research practices were changing. More scientists wanted to combine data collected at different parts of the electromagnetic spectrum. However, digital data recorded by scientists using a radio telescope in Australia was rarely compatible with that collected, for instance, by optical astronomers in California. "The data transport problem," a scientist at the Netherlands Foundation for Radio Astronomy noted, "is getting larger each year as more people seek to combine data from different instruments."[40] Moreover, each institution typically used its own software packages to read the often unique data formats its instruments produced. In other words, considerable data friction inhibited astronomers' ability to share research with colleagues or combine data collected at different instruments or telescopes. One way to grease this friction was to adopt a common format for astronomical data that the whole community used.

Starting in late 1976, a small group of astronomers at national observatories in the United States and Europe began to address the problem. At Kitt Peak National Observatory, for example, Donald C. Wells took a lead role. Wells started his career as a research astronomer but also taught himself how to program in FORTRAN and ALGOL.[41] After he moved to Kitt Peak's Tucson headquarters in 1972, Wells's interests shifted from astronomical research to information management and data handling. Because the national observatory's telescopes were accessible to any astronomer who successfully submitted a peer-reviewed proposal, Wells wanted to likewise build tools for data handling that a broader community could use.

In December 1976, Ronald Harten, an American-born radio astronomer working in the Netherlands, visited Wells at Kitt Peak. Radio astronomers, because their data is inherently "born electronic," had previously faced many of the challenges confronting their optical counterparts. Harten disliked the difficulty of moving data between radio telescopes in the Netherlands where it was collected and the offices where scientists later analyzed it. He told Wells about his experiments with a "magic record size" that might offer an initial step toward a solution.[42] At this point, different computer systems read data files in basic units of information interchange called "record lengths" that varied in size. If the chunk of data used was a common multiple of the various record

[37] Michael J. Disney, "Centre for Optical Data Analysis," unpublished report, n.d. (but likely 1978), RSE.

[38] Strom to Boyce, 6 July 1977 (cit. n. 11).

[39] "Minutes of the 1st Meeting, Panel on Astronomical Image and Data Processing," 27 October 1978, RSE. The overall situation closely resembles a "presumptive anomaly," as described by Edward Constant, *The Origins of the Turbojet Revolution* (Baltimore, 1980), 15.

[40] Ronald H. Harten to colleagues, 9 June 1978, DCW.

[41] Donald C. Wells, oral history interview by W. Patrick McCray, 16 July 2012.

[42] Donald C. Wells, "Happy Birthday, FITS!," 29 March 1992, e-mail message, http://www.cv.nrao.edu/fits/documents/overviews/history.news (accessed 15 October 2012).

lengths that commercially available computer systems could read, then this "universal commensurability" would enable the "packing and unpacking" of files on "a wide variety of computers."[43]

Wells and Harten devoted considerable time to engineering the "header" of the data record. Akin to what today is called "metadata," the header gives crucial information—where a picture was taken and with what instrument, celestial coordinates, observing conditions, and so forth—that precedes the data of the actual astronomical image. Because Wells and Harten represented the optical and radio astronomy communities, respectively, they needed to create headers general enough to apply to data collected in either wave band. They also wanted to create a header system that would be "flexible and self-defining" yet open to "indefinite expansion" in the future.[44]

As they developed their respective data interchange formats, neither was especially committed to the formats they had personally designed. As Wells wrote to Harten, "I believe that *any* well defined and widely accepted format is infinitely preferable to none." Harten agreed, noting that a "general purpose scheme" could attract the interest of as many scientists as possible. Through their respective efforts, Wells wrote, "the community is being exposed to our ideas," but the time was quickly coming for an "attempt to meld the opinions of a number of people to try to reach a compromise that can be accepted by all." Securing support from researchers at the major national observatories was highly desirable because, as Harten predicted, if scientists at these publicly funded institutions got on board, "then most of the battle is won."[45]

In January 1979, the National Science Foundation (NSF) arranged a meeting for representatives from the major national observatories in the United States to discuss digital image analysis. Given general agreement that a "tape interchange standard is important," a small committee representing Kitt Peak, the National Radio Astronomy Observatory, and NASA was set up to "facilitate the communication of digital data."[46] Three months later, Wells and his counterpart in the radio astronomy community, Eric Greisen, drafted an informal agreement based on a data format developed by Harten. The design for what they called the Flexible Image Transport System (FITS) would "implement the transfer of images between observatories" in a "general format" that was "flexible and contains virtually unlimited room for growth."[47] Their mutual acceptance of a common record length meant that data standardized into the FITS format could be read "on all computers commercially available in the U.S. today." Wells and Greisen successfully tested their system with a trial exchange of data, and the results were presented at an international meeting in June 1979.[48]

[43] D. C. Wells, E. W. Griesen, and R. H. Harten, "FITS: A Flexible Image Transport System," *Astron. Astrophys. Suppl. Ser.* 44 (1981): 363–70. In this era, an alphanumeric character was represented by a byte, which generally was 8 bits, where a bit is a single information element with a value of 0 or 1. The number ultimately chosen was 23,040 bits, equivalent to 2,880 8-bit bytes or 3,840 6-bit bytes. Moreover, it was evenly divisible by the byte lengths of computers on the market then; i.e., it is divisible by 6, 8, 12, 16, 18, 24, 32, 64, etc.

[44] Ibid.

[45] Letters between Wells and Harten, 17 May 1978 and 5 June 1978, DCW; emphasis in the original.

[46] The founding of the Space Telescope Science Institute (STScI), which manages the Hubble Space Telescope, was not formally announced until January 1981.

[47] "Draft of Flexible Image Transport System," 29 March 1979, DCW.

[48] Donald C. Wells, "FITS: A Flexible Image Transport System," in *International Workshop on Image Processing in Astronomy: Proceeding of the 5th Colloquium on Astrophysics, Trieste, June 4–8, 1979*, ed. G. Sedmak, M. Capaccioli, and R. J. Allen (Trieste, 1979), 445–71.

FITS offered astronomers a "syntax" for sharing data with each other or between their respective institutions. Greisen, Wells, and Harten saw that FITS also had value as an archival format. Their goal was that information preserved with FITS should be able to be read by all computer systems, old or new, in the future. Wells later claimed he saw this as analogous to James Madison's goal of protecting minority interests in the drafting of the U.S. Constitution.[49] Therefore, a policy of "once FITS, always FITS" was adopted to ensure backward compatibility. Moreover, aware of the potential value of astronomical data collected decades earlier in older formats such as photographic plates, they came to see FITS "not only as a way to talk to remote astronomers in the here and now" but also as a tool "to talk to future astronomers."[50]

Producing a common data exchange format, however, would be fairly worthless if other institutions didn't adopt it. This made promoting FITS a political as well as a technical activity. For Wells, this meant "trying to mobilize an opinion in the community of sharing data, of always using the same formats. I was trying to stamp out the heretics, people with alternative data formats."[51] To Wells's relief, astronomers quickly recognized the value of data standardization. By the end of 1980, national observatories in Sweden and Australia, in addition to those in the Netherlands and the United States, had adopted FITS as their basic data format.[52] In 1982, the International Astronomical Union officially sanctioned this by recommending that "all astronomical computer facilities recognize and support" FITS as the standard global interchange format for digital data.[53]

Of course, advocates of FITS could not compel astronomers to share their data. But for scientists inclined to do so, the process was now simpler and smoother. FITS presented astronomers with a lingua franca to foster easier sharing and archiving of digital data. Of course, the more scientists and institutions adopted FITS, the more essential it became for other scientists to enlist as well.[54] Once astronomers adopted FITS as an international data standard, it provided a potent oil to reduce friction inherent in the interinstitutional and transnational circulation of data. And, once digital data began to move and circulate more freely, some astronomers began to imagine a working world in which the digital tools to interact with it could also be shared.

[49] Donald Wells to W. Patrick McCray, e-mail message, 22 May 2011. It is worth noting that Wells sometimes used the famous quote from Benjamin Franklin to John Hancock—"We must indeed all hang together or, most assuredly, we shall all hang separately"—to encourage the astronomy community to unite behind FITS or another suitable standard.

[50] Donald C. Wells, "Speculations of the Future of FITS," in *Astronomical Data Analysis Software and Systems VI*, ed. Gareth Hunt and Harry Payne (San Francisco, 1997), 257–60. Lorraine Daston makes a similar point about collections created with the needs of future researchers in mind; see Daston, "The Sciences of the Archive," *Osiris* 27 (2012): 156–87.

[51] Donald C. Wells, oral history interview by W. Patrick McCray, 17 July 2012.

[52] See, e.g., Denis Warne of the Mount Stromlo and Siding Spring Observatories in Australia to Donald Wells, 2 August 1979, DCW.

[53] "Resolution C1," in *Proceedings of the 18th General Assembly, Transactions of the International Astronomical Union, Volume XVIIB*, ed. Richard M. West (Dordrecht, 1982), 46–7. One expert noted that, seen in the broader sense, FITS was perhaps the "first digital standard for data exchange," preceding formats like PDF, GIF, etc.; Robert Hanisch, personal communication with W. Patrick McCray, 1 September 2014.

[54] The process of getting other scientists to adopt FITS resembles, of course, the process of enlisting the participation of other actors described in Michel Callon and John Law, "On Interests and Their Transformation: Enrolment and Counter-Enrolment," *Soc. Stud. Sci.* 12 (1982): 615–25.

SHARING SCIENCE BY SHARING TOOLS

After retiring from Cardiff University, Michael Disney composed a roman à clef about his career in science. Originally trained as a theorist, in the 1970s Disney began to shift his research attention to observational astronomy. At the end of an observing run, astronomers like Disney would leave the telescope with "one or more inscrutable magnetic tapes" filled with data.[55] Before it could be "turned into useful astrophysics," this "crude and dirty" data had to be "calibrated, corrected, and cleaned."[56] Disney, according to his fictionalized account, found that "almost all of his time was going into writing and testing trivial but necessary computer programs . . . to carry out mundane but unavoidable housekeeping tasks."[57] Moreover, Disney observed that his colleagues were also writing their own algorithms and routines for processing data. "So," Disney (speaking via a fictional protagonist named "Cotteridge") asked, "why couldn't they share?"

One problem was technical. Before FITS became the community's data standard, different machines and programming languages created barriers to sharing. But, even with a common data format, astronomers faced a bewildering assortment of image processing programs. There was little in the way of standardization as researchers came up with fragmented solutions that were disorganized and ad hoc.[58] Software development efforts at one site were often duplicated at another place. Although astronomers often devoted considerable time to programming computers, Disney claimed they were "mostly incompetent" at this task, or they simply did not like doing it.[59] British astronomers described their situation as especially serious. The national investment was substantial, with the astronomy community receiving almost 20 percent of the research monies that came from the United Kingdom's Science Research Council. Data reduction, Disney and other scientists saw, was "beginning to create a bottleneck" where "a lot of valuable science" that their government had paid for "could be lost," placing them at a disadvantage compared to other members of their "highly competitive community."[60] Although building an adequate data processing infrastructure would not come cheaply, to "ignore or starve it" would "make as much sound sense as building a telescope in a cloudy site."[61]

The other hurdle was cultural. "Morgan," one of the scientists in Disney's fictional account, describes astronomers as "ambitious, competitive, selfish egoists. We all want to be the next Galileo, the next Isaac Newton." So why would a scientist with a better tool for data processing share it with a "more cunning rival who might use it to overtake us in the race for glory?" Cotteridge replies, "Because if we don't, we'll

[55] M. J. Disney and P. T. Wallace, "STARLINK," *Quart. J. Roy. Astron. Soc.* 23 (1982): 485–504.
[56] Ibid.
[57] Michael J. Disney, unpublished/untitled work of fiction, from chap. 4, "STARLINK, 1977"; electronic copy shared with W. Patrick McCray by Disney in 2013 (in McCray's working files).
[58] Richard S. Ellis, oral history interview with W. Patrick McCray, 28 November 2011.
[59] M. J. Disney, "Centre for Optical Data Analysis," unpublished report, February 1979, RSE.
[60] "Minutes of the 1st Meeting, Panel on Astronomical Image and Data Processing," (cit. n. 39); "Report of the Panel on Astronomical Image and Data Processing" (cit. n. 11), 6–7.
[61] "Report of the Panel on Astronomical Image and Data Processing" (cit. n. 11), 31. The reference to "cloudy site" was a barbed reference to the questionable decision made a decade earlier by British politicians to locate the 100-inch Isaac Newton Telescope at a poor location in the United Kingdom and a jab at the prioritizing of politics over scientific requirements; John Irvine and Ben Martin, "Assessing Basic Research: The Case of the Isaac Newton Telescope," *Soc. Stud. Sci.* 13 (1983): 49–86.

waste our entire lives writing trivial computer programs, leaving no time to do astronomy, vain-glorious or otherwise." Ah, true, Morgan retorts, but how do you "turn shits into saints?"[62] How, in other words, could one transform astronomy's moral economy in such a way that everyone is encouraged to give "at least a modest push to the common wheel"?[63]

In late 1978, with approval from the Science Research Council, a small Panel on Astronomical Image and Data Processing (PAIDP) chaired by Disney began to chart a new course.[64] Panel members perceived three basic options. One was to preserve the existing "laissez-faire" system. With "no central co-ordination," each institution would be "free to propose its own computer system and configuration" so as to "give astronomers the maximum choice."[65] However, this ran counter to the "spirit of co-operation" panel members wanted to foster.[66]

A distinct alternative to the existing system was to establish a single national center for all U.K. astronomers to use. This option offered streamlined management and funding as well as the elimination of redundancy as all data processing programs would be developed at a central facility. However, a single location meant that most users would have to travel to it, bypassing the "informally interactive system required by the very nature of the research."[67] One can see, in both cases, how perceptions of their discipline's moral economy were part of the committee's thinking.

In the best Goldilocks fashion, the PAIDP steered between these two extremes. It recommended a linked minicomputer network that the committee christened STARLINK.[68] Like other "star networks," STARLINK had a central node to which other networks' points would be linked. Besides acting as the switching center for communications, the central node would also service the network, update the computer systems, and make sure software was adequately documented. System software developed and shared between sites would ensure compatibility and prevent devolution back to the existing situation. Indeed, software sharing between scientists at different sites was the "laudable, and indeed compulsory goal" the PAIDP wished to achieve.[69]

In choosing the networked option, the PAIDP looked within its own national borders for an example. The Rutherford Laboratory, located near Oxford, hosted the Interactive Computing Facility. Set up in 1978, this facility was a general purpose computer network that scientists used to share software for applications like circuit design

[62] Disney, unpublished/untitled work of fiction, from chap. 4, "STARLINK, 1977" (cit. n. 57).

[63] Disney and Wallace, "STARLINK," (cit. n. 55), 501.

[64] The committee was chaired by Michael J. Disney. Joining him were three optical astronomers (Alec Boksenberg, Richard S. Ellis, and Robert Fosbury) as well as two computer experts (James Alty and Igor Alexsander).

[65] "Report of the Panel on Astronomical Image and Data Processing" (cit. n. 11), 13. Although no evidence exists to support this, one might speculate on the prevalence of laissez-faire in the context of British Thatcherite politics ca. 1980.

[66] "Discussion of Some Image Processing Alternatives for U.K. Astronomers," March 1979 Annex to "Minutes of the 5th Meeting of the Panel on Astronomical Image and Data Processing," 22 March 1979, RSE.

[67] "Report of the Panel on Astronomical Image and Data Processing" (cit. n. 11), 13.

[68] The recommendation is in ibid. A technical description of such a network is given in Lawrence G. Roberts and Barry D. Wessler, "Computer Network Development to Achieve Resource Sharing," in *AFIPS '70 (Spring) Proceedings of the May 5–7, 1970, Spring Joint Computer Conference*, ed. Harry L. Cooke (New York, 1970), 543–9.

[69] "Report of the Panel on Astronomical Image and Data Processing" (cit. n. 11), 10.

and fluid mechanics.[70] The PAIDP's recommendation to base STARLINK at Rutherford, rather than at one of the royal observatories or at a university with a large and/or eminent astronomy department, reflected the priority given to software sharing over astronomical research per se. At the same time, the PAIDP insisted that "STARLINK must at all times respond to astronomical needs."[71] However, there was an obvious tension between these two goals that became more pronounced over time.

In its initial configuration, STARLINK was based on six VAX-11/780 minicomputer machines. Besides the central node in Chilton, other machines went to places such as Cambridge, University College London, and the Royal Greenwich Observatory, choices made based on estimates that 80 percent of U.K. astronomers worked at or within twenty miles of these sites. Leased telephone lines from Britain's Post Office connected the nodes. At each of STARLINK's sites, astronomers could access two image-display systems that allowed them to interact with their data in real time. The STARLINK project also adopted FITS as its data interchange format. The whole system—with an initial cost of £1.8 million—was inaugurated in October 1980.[72]

Implementing STARLINK as a tool for sharing software proved more difficult than astronomers originally expected. There were several reasons for this mismatch between aspiration and actualization. First, the science community expressed "conflicting requirements" as to how STARLINK should function. On the one hand, scientists wanted the "immediate no-nonsense development of a large number of application programs" they could access via STARLINK.[73] These "ultra-pragmatists" believed that "little if any supporting software over and above application routines were needed."[74] They were opposed by "idealists" who "cannot brook the slightest departure from complete portability across machine types."[75] In short, some scientists wanted to start using STARLINK in a quick and dirty fashion while others wanted to wait for a "comprehensive, soundly architectured, easy to use, and efficient" system. A fault line also ran between professional disciplines. Astronomers, Disney observed, wanted immediate results "no matter how inefficient, *ad hoc*, and inelegant" the data processing techniques were that yielded them. Computer scientists, in contrast, were as much "concerned with methods as with a particular astronomer's results."[76]

STARLINK advocates also had to contend with what Disney's fictional character Morgan had indecorously called the "saints versus shits" problem. Even before STARLINK was officially launched, astronomers anticipated that scientists at the various network nodes might "implement only software that was locally in demand."[77]

[70] Information on the Interactive Computing Facility is at http://www.chilton-computing.org.uk/acd/literature/annual_reports/p004.htm (accessed 10 July 2014). In 1979, the Appleton Laboratory moved to Chilton and was combined with the Rutherford Laboratory to create the Rutherford Appleton Laboratory.

[71] "Report of the Panel on Astronomical Image and Data Processing" (cit. n. 11), 16, 20.

[72] The inauguration featured a computer program that allowed astronomical images to be retrieved over the network from the various STARLINK nodes and then assembled on an image display terminal at Chilton. Described in "Inauguration of STARLINK," *Enterprise: STARLINK Information Bulletin*, no. 3 (November 1980): 1. See also http://www.chilton-computing.org.uk/acd/starlink/p004.htm (accessed 29 December 2012).

[73] "STARLINK Applications Software," *Enterprise: STARLINK Information Bulletin*, no. 3 (November 1980): 2.

[74] Disney and Wallace, "STARLINK" (cit. n. 55), 493.

[75] Ibid.

[76] Mike Disney, "Concluding Remarks," in *International Workshop on Image Processing in Astronomy*, ed. G. Sedmak, M. Capaccioli, and R. J. Allen (Trieste, 1979), 495–500, on 498.

[77] "Discussion of Some Image Processing Alternatives" (cit. n. 66).

This, of course, ran counter to STARLINK's original purpose—avoiding wasted effort inherent in the original laissez-faire model by "coordinating much of the software common to many requirements into a few universal and centrally supported packages."[78] Even more worrisome was the possibility that some centers might develop especially innovative software and "be reluctant to share it."[79] STARLINK's premise was that users would develop data processing software and share it over the network. However, once an institution received its VAX machine and interactive terminals, there was no way to compel its astronomers to share programs. In fact, astronomers actually had incentive not to share software via STARLINK.

Consider the differences between FITS and STARLINK. All scientists worked with data; FITS ensured its portability between researchers and institutions, benefiting all. Researchers interacted with STARLINK, however, after they already had their rough data. At this point, the impetus was to convert data into career-enhancing scientific results and publications. If one already had superior image processing software—perhaps written personally or by a colleague—then there might be reduced incentive to share it via STARLINK. Soon after the system was inaugurated, Disney restated his belief that there was "*no alternative* to sharing software development." Doing so, however, demanded "alertness, openness, generosity, and a strong spirit of compromise."[80] Fortunately, STARLINK could draw on an expanding user community, which included enough people possessing a combination of altruism and self-interest—there were over 1,000 users by 1988—to help oil Disney's "common wheel."

While not wanting to oversimplify, one can situate the aspirations Disney and other advocates had for STARLINK and other community-developed data processing tools in a broader context. Although it was years before the open-source software movement, there was an ethos of sharing in the 1970s-era computer culture, typified by the members of the Homebrew Computer Club and other hobbyist groups.[81] Of course, the groovy world of Bay Area hackers was considerably removed from research programs at Cambridge or Durham. Nonetheless, one can detect a common focus on sharing that reflects larger community aspirations and norms.

Likewise, Disney and other STARLINK advocates expressed a certain sense of idealism, perhaps even technological utopianism, about what their digital systems might accomplish. As Disney later recalled in his fictionalized account, Cotteridge "was designing The Future." Likewise, Disney and his colleagues speculated on how systems like STARLINK might affect the "shape of astronomy in the 21st century."[82] To look forward, Disney looked back in time, noting how the "cheap plane ticket" had caused the "backyard telescope and the staff astronomer" to give way to the "remote National facility and the guest observer."[83] Software systems and data sets accessed via high-

[78] R. J. Dickens, R. A. E. Fosbury, and P. T. Wallace, "The Development of STARLINK Applications Software," unpublished report, 6 October 1980, RSE. Early ARPANET developers faced similar issues as the reality of resource sharing often fell below expectations; Janet Abbate, *Inventing the Internet* (Cambridge, Mass., 1999), 96–7.

[79] "Discussion of Some Image Processing Alternatives" (cit. n. 66).

[80] Disney and Wallace, "STARLINK" (cit. n. 55), 501; emphasis in the original.

[81] Steven Levy, *Hackers: Heroes of the Computer Revolution* (New York, 1984); Christopher M. Kelty, *Two Bits: The Cultural Significance of Free Software* (Durham, N.C., 2008).

[82] I. Elliott, "Starlink," *Irish Astron. J.* 14 (1980): 197; this was elaborated on in Disney and Wallace, "STARLINK" (cit. n. 55).

[83] Disney and Wallace, "STARLINK" (cit. n. 55), 503.

bandwidth links would cause similar changes in astronomy's working world. In time, "authors on different continents could collaborate," and eventually, the "community of locations" would yield to the "international team based on community of interest."[84] And, as the network grew, Disney imagined using systems like STARLINK to share not only data processing tools but "equations, drawings, papers, data, and pictures." It will be, he predicted, as if all astronomers "live in the same electronic corridor," where data circulated in a more frictionless fashion.[85]

By 1990, much of the technical friction in the way of achieving this ideal had been smoothed over. Data were almost always "born digital," while older photographic forms of data could be routinely converted to zeros and ones. The astronomical community had agreed-upon data standards, and it could share tools for processing data. Nonetheless, at least one major barrier remained to achieving Disney's dream. A National Academy of Sciences report described the situation in the United States thus: "most ground-based astronomical data obtained outside the national observatories are treated as the private property of the observer," and, consequently, there was "no imperative" to share it.[86] Overcoming this particular point of friction—part technical and part social—would require significant additional changes in how astronomers thought about intellectual property, sharing, proprietary rights, and so forth—that is, the norms and behaviors underpinning the community's moral economy.

ARCHIVING THE UNIVERSE

In 2003, information managers at the Space Telescope Science Institute reviewing data usage statistics noticed something interesting. The total number of publications astronomers produced using the HST the previous year was just under 600. For the first time, half of these referred papers used archived Hubble data, a percentage that continued to climb.[87] Even more notable was the fact that roughly 40 percent of HST publications had used only archived data.

Modern astronomy's emergence as an archival science can be detected in other ways. Starting in the mid-1960s, the astronomy community regularly conducted a disciplinary review. These "decadal reports" set national research priorities for the next ten years. It was not until the third such survey, concluded in 1980, that the word "data" appeared in association with "archives."[88] The online SAO/NASA Astrophysics Data System tells a similar story. Searches of papers published between 1970 and 1980 show that the phrase "digital archive" was not used at all and that "data archive" appeared in paper titles on only four occasions. Jumping ahead a decade, one finds nineteen titles containing "digital archive" and 244 instances of "data archive." The

[84] Ibid.
[85] Ibid. Historians, of course, will recognize the familiar trope here as new communication tools were imbued with all sorts of utopian desires, including that of the "paperless office." See, e.g., "Office of the Future," *Business Week*, 30 June 1975, 48–70.
[86] John N. Bahcall, ed., *The Decade of Discovery in Astronomy and Astrophysics* (Washington, D.C., 1991), 96.
[87] This statement is based on statistics at http://archive.stsci.edu/hst/bibliography/pubstat.html (accessed 10 July 2014). I am using HST as an example because, among major public observatories, it maintains the most accessible records of this sort. Similar patterns would be found at other observatories.
[88] George Field, ed., *Astronomy and Astrophysics for the 1980's: Reports of the Panels* (Washington, D.C., 1983).

word "archive" alone shows a more dramatic jump, from thirty-one instances to almost 800. Such numbers confirm that the data archive—once containing tangible photographic plates but now composed of digital bits and pixels—had become an important research site.

Two developments helped catalyze this shift. In the 1990s, the astronomy community carried out several large survey projects in a variety of wavelengths. Often conducted with specially designed instruments, these surveys were planned from the outset to produce coherent data sets recorded with well-defined standards. The flurry in survey activity coincided with changes in computing technology and astronomical instrumentation. Computer processing speeds and memory storage capacity continued to march in step with Moore's law. While it took a quarter century for the light-collecting area of telescopes to double, the number of pixels in astronomers' digital detectors doubled every few years. Just as critical was the emergence of the Internet and the World Wide Web as legitimate research tools. The zeitgeist of the dot-com boom and the desire of astronomers to take advantage of new digital tools merged, for instance, when Johns Hopkins University astronomer Alexander Szalay and Microsoft computer scientist Jim Gray described astronomers' plans to "make the Internet act as the world's best telescope—a World-Wide Telescope."[89]

As astronomers in the United States and Europe began to consider building so-called virtual observatories—the term first appears in the literature around 1997—they also reevaluated the fundamental nature of their data. As Szalay and Gray, two of the most prominent advocates for virtual observatories, described it, astronomers' raw digital data was a complex assortment of "fluxes . . . spectra . . . individual photon events." Moreover, unlike data in other disciplines that "can be frozen and distributed to other locations," astronomical data often needs reprocessing and recalibration such that it "stays 'live' much longer . . . [and] needs an active 'curation.'" However, because each research group had its "own historical reasons" and methods for saving its data "one way or another," a single centralized repository, like the molecular biologists' GenBank, didn't seem feasible. They concluded that a "federated" system that would unite existing databases seemed more realizable.[90]

From the inception of the idea, rhetoric around virtual observatories was imbued with utopian aspirations that accompanied other Internet-related endeavors at the turn of the century. Virtual observatory advocates gushed about the liberating possibilities of "mining the sky." For the "clever people who don't have access to a big telescope," said Caltech's George Djorgovski, a virtual observatory "will allow them to do first-rate observational astronomy."[91] Claims made on both sides of the Atlantic hinted at the possibility of political, not just scientific, revolution. Virtual observatories could "lead to a true democratization of astronomy" and represented a "fresh wind blowing through the graveyard of old and unused data."[92] Riffing on journalist Thomas Friedman's best-selling 2005 book *The World Is Flat*, Matt Mountain suggested that

[89] Alexander Szalay and Jim Gray, "The World-Wide Telescope," *Science* 293 (2001): 2037–40.
[90] Ibid., 2038.
[91] Ron Cowen, "Mining the Sky: Taking Some Big Bytes of the Universe," *Sci. News* 159 (2001): 124–5.
[92] "Democratization" quote from Caltech's Robert Brunner in Govert Schilling, "The Virtual Observatory Moves Closer to Reality," *Science* 289 (2000): 238–9; "graveyard" quote from German astronomer Wolfgang Voges in Toni Feder, "Astronomers Envision Linking World Data Archives," *Phys. Today* 55 (2002): 20–2, on 20.

publicly accessible astronomical databases might "make the sky flat" by increasing access to data resources.[93] Not everyone was so sanguine, however. One astronomer predicted that virtual observatories only "breed a generation of astronomers who sift through data without knowing about instruments."[94] Others questioned data archives' costs, technical challenges, and perceived banality compared to building a giant new telescope.

Astronomers' imaginings of virtual observatories were more far-reaching in scope than other data-focused efforts described in this essay. Kibblewhite's APM and STARLINK were locally bound machines or systems. FITS gradually became a community standard and a necessary first step in facilitating digital data sharing, but it addressed a specific technical problem. Ambitions for virtual observatories, in contrast, were fully transnational in scope, bringing together databases from disparate countries and observations made across the spectrum so that they might constitute "one uniform, consistent data set."[95]

In the United States, a blue-ribbon panel of astronomers gathered by the National Academy of Sciences nudged the "National Virtual Observatory" (NVO) forward by making it a top priority for the early twenty-first century.[96] Researchers from observatories and computer science centers put together an implementation plan, and the NSF awarded $10 million toward their efforts.[97] In the United States, proponents described it as a "new research environment for astronomy with massive datasets," the creation of which would be "technology-enabled but science-driven."[98] Similar efforts emerged in the United Kingdom and the European Union, resulting in the creation of the International Virtual Observatory Alliance. In 2010, the NVO transitioned to become the Virtual Astronomical Observatory, while Elsevier's launch of the journal *Astronomy and Computing* provided a forum for peer-reviewed publications. By 2014, when the effort concluded, some $16 million from the NSF and NASA had helped dozens of scientists and computer programmers begin to build a more robust data infrastructure for astronomy.[99]

Digital archives retained the potential for slowly eroding established data-sharing conventions. Many of these changes accelerated as more scientists embraced a multiwavelength astronomy that relied on data collected at multiple facilities. Taxpayer-funded "Big Science" has also helped to drive the process. In August 1982, for example, Riccardo Giacconi, director of the recently formed Space Telescope Science Institute, which would handle science operations for the as-yet-to-be-launched HST, explained how individual researchers' data would have a proprietary period of just

[93] Matt Mountain, "Flattening the Astronomy World," *Phys. Today* 67 (2014): 8–10.
[94] Feder, "Astronomers" (cit. n. 92).
[95] Jim Gray in Cowen, "Mining the Sky" (cit. n. 91).
[96] Christopher McKee and Joseph Taylor, eds., *Astronomy and Astrophysics in the New Millennium* (Washington, D.C., 2000); the decadal report recommended that $60 million be directed to the NVO.
[97] "Building the Framework for the National Virtual Observatory," 23 April 2001 proposal to the National Science Foundation, electronic copy archived at http://www.us-vo.org/pubs/index.cfm (accessed 30 June 2014).
[98] "Towards a National Virtual Observatory: Science Goals, Technical Challenges, and Implementation Plan," white paper prepared for National Academy of Sciences, 8 June 2000, electronic copy archived at http://www.us-vo.org/pubs/index.cfm (accessed 30 June 2014).
[99] Robert Hanisch, "A Brief History of the US VO Effort," 10 July 2014 presentation, Pasadena, Calif.; copies of slides in W. Patrick McCray's working papers. Also, Robert Hanisch, personal communication with McCray, 8 September 2014.

one year. After that, the data would "be made available to the community at large," becoming a public good.[100]

Jump ahead thirteen years. For ten consecutive days in late 1995, Hubble took 342 exposures of a small region of the sky in the constellation Ursa Major. The resulting data, once released to astronomers, caused great public wonderment.[101] Since then, the Hubble Deep Field has become an iconic scientific image. The thousands of jewel-like galaxies it revealed have been reproduced on calendars, coffee cups, and screen savers, while scientists used the shared data set to produce hundreds of refereed papers.[102] Far from being the property of a single investigator, the Hubble Deep Field (and similar surveys that followed) presented astronomical data as a shared community resource. One wonders what Disney's dyspeptic character Morgan might have thought.

* * *

In 1945, Henry Norris Russell advised a colleague about to assume the directorship of a major observatory that he should hire a "good man who knows modern electronic instrumentation."[103] The subsequent postwar "de-astronomization of astronomy" brought many new experts into the field.[104] This trend has continued with a redefinition of who and what an astronomer is. In the decades after Russell's suggestion, astronomers migrated away from their analog data traditions, taking first an electronic and then a digital turn. Events of the past decade suggest another turn in process. Starting in 2008, a new term—"astroinformatics"—began to appear in the international online repository of astronomy papers. Proponents described it as a "new data-oriented paradigm for astronomy."[105] The neologism reflects the tendency, seen in many scientific fields, to embrace, or at least come to terms with, a broader data turn—that is, bio-

[100] Riccardo Giacconi, "Science Operations with Space Telescope (NASA CP-2244)," in *The Space Telescope Observatory*, ed. Donald N. B. Hall (NASA, 1982), 1–15, on 11. An earlier statement of this policy can be found at least a year earlier in NASA technical documents such as "Statement of Work for the Space Telescope Science Institute (STScI)," 15 April 1981, 8; copies of both documents are in the library at the STScI in Baltimore.

[101] NASA's original press release is available at http://hubblesite.org/newscenter/archive/releases/1996/01/text/ (accessed 1 July 2014), while technical details are at http://www.stsci.edu/ftp/science/hdf/hdf.html (accessed 1 July 2014) and http://www.stsci.edu/stsci/meetings/shst2/williamsr.html (accessed 1 July 2014). See also Elizabeth A. Kessler, *Picturing the Cosmos: Hubble Space Telescope Images and the Astronomical Sublime* (Minneapolis, 2012).

[102] The original Hubble Deep Field paper—Robert E. Williams, Brett Blacker, Mark Dickinson, W. Van Dyke Dixon, Henry C. Ferguson, Andrew S. Fruchter, Mauro Giavalisco, et al., "The Hubble Deep Field: Observations, Data Reduction, and Galaxy Photometry," *Astron. J.* 112 (1996): 1335–89—has been, according to Google Scholar (7 August 2016), cited more than 1,300 times. STScI director Williams even noted that some requests for HST time offered shorter proprietary data periods to get "an edge" vis-à-vis more stingy proposals; Toni Feder, "Space Scientists Split of Proprietary Data Rights," *Phys. Today* 51 (1998): 52–3.

[103] Henry Norris Russell to Ira S. Bowen, 3 November 1945, cited in David DeVorkin, "Electronics in Astronomy: Early Applications of the Photoelectric Cell and Photomultiplier for Studies of Point-Source Celestial Phenomena," *Proc. IEEE* 73 (1985): 1205–20, on 1220.

[104] Jesse L. Greenstein, oral history interview with Rachel Prud'homme, 16 March 1982, Center for History of Physics, American Institute of Physics, College Park, Md.

[105] Quote from https://asaip.psu.edu/Articles/astroinformatics-in-a-nutshell (accessed 15 June 2015). See also Kirk Borne, Alberto Accomazzi, Joshua Bloom, Robert Brunner, Douglas Burke, Nathaniel Butler, David F. Chernoff, et al., "Astroinformatics: A 21st Century Approach to Astronomy," *Astronomy and Astrophysics Decadal Survey, Position Papers*, no. 6 (2009), http://adsabs.harvard.edu/abs/2009astro2010P...6B (accessed 15 June 2015).

informatics, genomics, proteomics, and so forth.[106] Whether astroinformatics coalesces into a sustainable community is too early to tell. But one outcome is already clear: contemporary astronomers need no longer be producers of scientific data but can instead make careers as consumers of it.

Looking across astronomy's three overlapping data eras—analog, electronic, and digital—one can make a broader observation about data friction and the norms and practices about data sharing. In the traditional photographic era, data friction was high. It was difficult and expensive to trade, share, move, and reproduce raw data. Yet the "rules" of the community's moral economy were relatively simple—data almost always belonged to the individual who collected it; what circulated was mostly processed data and findings. As astronomers entered the born-digital era, new hardware and software substantially reduced data friction, making it easier to move and share information in the form of raw data. But, at this point, navigating astronomy's moral economy also became harder as issues around data sharing, ownership, and access became more complex.[107]

Astronomers' experiences navigating their community's norms and expectations about data sharing invite comparisons to those of scientists from other disciplines. A tempting topic for future research, a full exegesis is not possible here. However, a few observations can be made. Kohler's detailed exploration of the drosophilists' network of sharing and exchange, for example, highlights the importance of free exchange of fruit fly stocks, the role of "enlightened self-interest" in promoting this exchange, and the "unspoken rules of etiquette" that governed this circulation.[108] One factor that facilitated the acceptance of these rules, he suggests, was the relative abundance of *Drosophila* as a research material, a situation made possible by large-scale breeding and stock keeping.

In comparison, astronomers' perception that the amount of data available to them was expanding at a rate that seemed overwhelming did not, by itself, reshape their moral economy. Instead, technological interventions in the form of data standards and new data-handling tools were required. An abundance of research material alone—in this case, the rapidly rising flood of data—did not, *pace* Kohler, suddenly alter astronomers' long-held if often unspoken rules and expectations about sharing it.[109]

In this essay, I have discussed representative examples of astronomers' engagement with their data: the development of machines in the early 1970s to convert analog data to digital format, computer-savvy astronomers working at national observatories in the United States and Europe in the late 1970s to create a community-wide data standard to facilitate sharing, attempts in the United Kingdom circa 1980 to share software tools for data processing, and early twenty-first-century international efforts to build

[106] Previously, other fields have also been portrayed as extensions of information science; see Timothy Lenoir, "Shaping Biomedicine as an Information Science," in *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*, ed. Mary Ellen Bowden, Trudi Bellardo Hahn, and Robert V. Williams (Medford, N.J., 1999), 27–45; David Baltimore, "How Biology Became an Information Science," in *The Invisible Future: The Seamless Integration of Technology into Everyday Life*, ed. Peter J. Denning (New York, 2001), 43–55.

[107] My thanks to David C. Brock for valuable comments on this point.

[108] Kohler, *Lords of the Fly* (cit. n. 15).

[109] My thanks to the perceptive reviewer who suggested this as a point of comparison. The effect of data's "bigness" is a point taken up more recently in Sabina Leonelli, "What Difference Does Quantity Make? On the Epistemology of Big Data in Biology," *Big Data & Soc.* 1 (2014): 1–11.

publicly accessible online data archives. From these vignettes, a more general typology emerges of astronomers' data-driven working world beyond the telescope.

Data conversion. Astronomers saw value in transforming data recorded on traditional photographic plates into digital data that they could then interact with. Subsequently, researchers constructed increasingly sophisticated systems to render analog data into a more pliable digital format.

Data standardization. In all technological systems, agreed-upon standards serve as political as well as technical tools—the means to discipline unruly arrays of measurements, signals, and so forth, as well as the communities associated with them.[110] Likewise, in astronomy, seemingly mundane standards served social and technical purposes by encouraging additional order and rationality.

Data processing. Once "raw" data had been collected, a key task for the astronomer was processing it. This general term encompasses a wide range of specific actions, but, in all cases, the goal was to interact with it so as to produce meaningful scientific information.

Data archiving and access. Traditionally, optical astronomers imagined their data as personal property. In some cases, observatories maintained physical libraries of data in the form of photographic plates to which staff contributed. However, as scientists gained access to increasingly expensive facilities, particularly those funded with public money, the idea of data as a public good became more powerful and widespread. Subsequently, some astronomers began to envision a more seamless system in which data collected by many different telescopes are managed and stored.

All of these activities are embedded in a larger framework of circulation and sharing and conditioned by astronomy's political and moral economies. To be sure, while new tools and new technologies can help reduce data friction, they cannot by themselves eliminate its accompanying social friction. To paraphrase Shakespeare, that particular fault lies not in the stars but in ourselves.

[110] Andrew L. Russell, "Standardization in History: A Review Essay with an Eye to the Future," in *The Standards Edge: Future Generations*, ed. Sherrie Bolin (Ann Arbor, Mich., 2005), 247–60.