



I Have Seen the Paradigm Shift, and It Is Us

JOHN WILBANKS
Creative Commons

TEND TO GET NERVOUS WHEN I HEAR TALK OF PARADIGM SHIFTS. The term itself has been debased through inaccurate popular use—even turning into a joke on *The Simpsons*—but its original role in Thomas Kuhn’s *Structure of Scientific Revolutions* [1] is worth revisiting as we examine the idea of a Fourth Paradigm and its impact on scholarly communication [2].

Kuhn’s model describes a world of science in which a set of ideas becomes dominant and entrenched, creating a worldview (the infamous “paradigm”) that itself gains strength and power. This set of ideas becomes powerful because it represents a plausible explanation for observed phenomena. Thus we get the luminiferous aether, the miasma theory of infectious disease, and the idea that the sun revolves around the Earth. The set of ideas, the worldview, the paradigm, gains strength through incrementalism. Each individual scientist tends to work in a manner that adds, bit by bit, to the paradigm. The individual who can make a big addition to the worldview gains authority, research contracts, awards and prizes, and seats on boards of directors.

All involved gain an investment in the set of ideas that goes beyond the ideas themselves. Industries and governments (and the people who work in them) build businesses and policies that depend on the worldview. This adds a layer of defense—an immune system of sorts—that protects the worldview against attack.



Naysayers are marginalized. New ideas lie fallow, unfunded, and unstaffed. Fear, uncertainty, and doubt color perceptions of new ideas, methods, models, and approaches that challenge the established paradigm.

Yet worldviews fall and paradigms shatter when they stop explaining the observed phenomena or when an experiment conclusively proves the paradigm wrong. The aether was conclusively disproven after hundreds of years of incrementalism. As was miasma, as was geocentrism. The time for a shift comes when the old ways of explaining things simply can no longer match the new realities.

This strikes me as being the idea behind Jim Gray's argument about the fourth data paradigm [3] and the framing of the "data deluge"—that our capacity to measure, store, analyze, and visualize data is the new reality to which science must adapt. Data is at the heart of this new paradigm, and it sits alongside empiricism, theory, and simulation, which together form the continuum we think of as the modern scientific method.

But I come to celebrate the first three paradigms, not to bury them. Empiricism and theory got us a long way, from a view of the world that had the sun revolving around the Earth to quantum physics. Simulation is at the core of so much contemporary science, from anthropological re-creations of ancient Rome to weather prediction. The accuracy of simulations and predictions represents the white-hot center of policy debates about economics and climate change. And it's vital to note that empiricism and theory are essential to a good simulation. I can encode a lovely simulation on my screen in which there is no theory of gravity, but if I attempt to drive my car off a cliff, empiricism is going to bite my backside on the way down.

Thus, this is actually not a paradigm shift in the Kuhnian sense. Data is not sweeping away the old reality. Data is simply placing a set of burdens on the methodologies and social habits we use to deal with and communicate our empiricism and our theory, on the robustness and complexity of our simulations, and on the way we expose, transmit, and integrate our knowledge.

What needs to change is our paradigm of ourselves as scientists—not the old paradigms of discovery. When we started to realize that stuff was made of atoms, that we were made of genes, that the Earth revolved around the sun, those were paradigm shifts in the Kuhnian sense. What we're talking about here cuts across those classes of shift. Data-intensive science, if done right, will mean more paradigm shifts of scientific theory, happening faster, because we can rapidly assess our worldview against the "objective reality" we can so powerfully measure.

The data deluge strategy might be better informed by networks than by Kuhnian



dynamics. Networks have a capacity to scale that is useful in our management of the data overload—they can convert massive amounts of information into a good thing so the information is no longer a “problem” that must be “solved.” And there is a lesson in the way networks are designed that can help us in exploring the data deluge: if we are to manage the data deluge, we need an open strategy that follows the network experience.

By this I mean the “end-to-end,” layer-by-layer, designed information technology and communications networks that are composed of no more than a stack of protocols. The Internet and the Web have been built from documents that propose standard methods for transferring information, describing how to display that information, and assigning names to computers and documents. Because we all agree to use those methods, because those methods can be used by anyone without asking for permission, the network emerges and scales.

In this view, data is not a “fourth paradigm” but a “fourth network layer” (atop Ethernet, TCP/IP, and the Web [4]) that interoperates, top to bottom, with the other layers. I believe this view captures the nature of the scientific method a little better than the concept of the paradigm shift, with its destructive nature. Data is the result of incremental advances in empiricism-serving technology. It informs theory, it drives and validates simulations, and it is served best by two-way, standard communication with those layers of the knowledge network.

To state it baldly, the paradigm that needs destruction is the idea that we as scientists exist as un-networked individuals. Now, if this metaphor is acceptable, it holds two lessons for us as we contemplate network design for scholarly communication at the data-intensive layer.

The first lesson, captured perfectly by David Isenberg, is that the Internet “derives its disruptive quality from a very special property: IT IS PUBLIC.” [5] It’s public in several ways. The standard specifications that define the Internet are themselves open and public—free to read, download, copy, and make derivatives from. They’re open in a copyright sense. Those specifications can be adopted by anyone who wants to make improvements and extensions, but their value comes from the fact that a lot of people use them, not because of private improvements. As Isenberg notes, this allows a set of “miracles” to emerge: the network grows without a master, lets us innovate without asking for permission, and grows and discovers markets (think e-mail, instant messaging, social networks, and even pornography). Changing the public nature of the Internet threatens its very existence. This is not intuitive to those of us raised in a world of rivalrous economic goods and



traditional economic theory. It makes no sense that Wikipedia exists, let alone that it kicks Encyclopedia Britannica to the curb.

As Galileo might have said, however, “And yet it moves.” [6] Wikipedia does exist, and the network—a consensual hallucination defined by a set of dry requests for comments—carries Skype video calls for free between me and my family in Brazil. It is an engine for innovation the likes of which we have never seen. And from the network, we can draw the lesson that new layers of the network related to data should encode the idea of publicness—of standards that allow us to work together openly and transfer the network effects we know so well from the giant collection of documents that is the Web to the giant collections of data we can so easily compile.

The second lesson comes from another open world, that of open source software. Software built on the model of distributed, small contributions joined together through technical and legal standardization was another theoretical impossibility subjected to a true Kuhnian paradigm shift by the reality of the Internet. The ubiquitous ability to communicate, combined with the low cost of acquiring programming tools and the visionary application of public copyright licenses, had the strangest impact: it created software that worked, and scaled. The key lesson is that we can harness the power of millions of minds if we standardize, and the products can in many cases outperform those built in traditional, centralized environments. (A good example is the Apache Web server, which has been the most popular Web server software on the Internet since 1996.)

Creative Commons applied these lessons to licensing and created a set of standard licenses for cultural works. These have in turn exploded to cover hundreds of millions of digital objects on the network. Open licensing turns out to have remarkable benefits—it allows for the kind of interoperability (and near-zero transaction costs) that we know from technical networks to occur on a massive scale for rights associated with digital objects such as songs and photographs—and scientific information.

Incentives are the confounding part of all of this to traditional economic theory. Again, this is a place where a Kuhnian paradigm shift is indeed happening—the old theory could not contemplate a world in which people did work for free, but the new reality proves that it happens. Eben Moglen provocatively wrote in 1999 that collaboration on the Internet is akin to electrical induction—an emergent property of the network unrelated to the incentives of any individual contributor. We should not ask why there is an incentive for collaborative software development any more than we ask why electrons move in a current across a wire. We should instead ask,



what is the resistance in the wire, or in the network, to the emergent property? Moglen's Metaphorical Corollaries to Faraday's Law and Ohm's Law¹ still resonate 10 years on.

There is a lot of resistance in the network to a data-intensive layer. And it's actually not based nearly as much on intellectual property issues as it was on software (although the field strength of copyright in resisting the transformation of peer-reviewed literature is very strong and is actively preventing the "Web revolution" in that realm of scholarly communication). With data, problems are caused by copyright,² but resistance also comes from many other sources: it's hard to annotate and reuse data, it's hard to send massive data files around, it's hard to combine data that was not generated for recombination, and on and on. Thus, to those who didn't generate it, data has a very short half-life. This resistance originates with the paradigm of ourselves as individual scientists, not the paradigms of empiricism, theory, or simulation.

I therefore propose that our focus be Moglen-inspired and that we resist the resistance. We need investment in annotation and curation, in capacity to store and render data, and in shared visualization and analytics. We need open standards for sharing and exposing data. We need the RFCs (Requests for Comments) of the data layer. And, above all, we need to teach scientists and scholars to work in this new layer of data. As long as we practice a micro-specialization guild culture of training, the social structure of science will continue to provide significant resistance to the data layer.

We need to think of ourselves as connected nodes that need to pass data, test theories, access each others' simulations. And given that every graph about data collection capacity is screaming up exponentially, we need scale in our capacity to use that data, and we need it badly. We need to network ourselves and our knowledge. Nothing else we have designed to date as humans has proven to scale as fast as an open network.

Like all metaphors, the network one has its limits. Networking knowledge is harder than networking documents. Emergent collaboration in software is easier

¹ "Moglen's Metaphorical Corollary to Faraday's Law says that if you wrap the Internet around every person on the planet and spin the planet, software flows in the network. It's an emergent property of connected human minds that they create things for one another's pleasure and to conquer their uneasy sense of being too alone. The only question to ask is, what's the resistance of the network? Moglen's Metaphorical Corollary to Ohm's Law states that the resistance of the network is directly proportional to the field strength of the 'intellectual property' system." [7]

² Data receives wildly different copyright treatment across the world, which causes confusion and makes international licensing schemes complex and difficult. [8]



because the tools are cheap and ubiquitous—that’s not the case in high-throughput physics or molecular biology. Some of the things that make the Web great don’t work so well for science and scholarship because the concept of agreement-based ratings find you only the stuff that represents a boring consensus and not the interesting stuff along the edges.

But there is precious little in terms of alternatives to the network approach. The data deluge is real, and it’s not slowing down. We can measure more, faster, than ever before. We can do so in massively parallel fashion. And our brain capacity is pretty well frozen at one brain per person. We have to work together if we’re going to keep up, and networks are the best collaborative tool we’ve ever built as a culture. And that means we need to make our data approach just as open as the protocols that connect computers and documents. It’s the only way we can get the level of scale that we need.

There is another nice benefit to this open approach. We have our worldviews and paradigms, our opinions and our arguments. It’s our nature to think we’re right. But we might be wrong, and we are most definitely not completely right. Encoding our current worldviews in an open system would mean that those who come along later can build on top of us, just as we build on empiricism and theory and simulation, whereas encoding ourselves in a closed system would mean that what we build will have to be destroyed to be improved. An open data layer to the network would be a fine gift to the scientists who follow us into the next paradigm—a grace note of good design that will be remembered as a building block for the next evolution of the scientific method.

REFERENCES

- [1] T. S. Kuhn, *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1996.
- [2] G. Bell, T. Hey, and A. Szalay, “Beyond the Data Deluge,” *Science*, vol. 323, pp. 1297–1298, Mar. 6, 2009, doi: 10.1126/science.1170411.
- [3] J. Gray and A. Szalay, “eScience - A Transformed Scientific Method,” presentation to the Computer Science and Technology Board of the National Research Council, Mountain View, CA, Jan. 11, 2007. (Edited transcript in this volume.)
- [4] Joi Ito, keynote presentation at ETech, San Jose, CA, Mar. 11, 2009.
- [5] “Broadband without Internet ain’t worth squat,” by David Isenberg, keynote address delivered at Broadband Properties Summit, accessed on Apr. 30, 2009, at <http://isen.com/blog/2009/04/broadband-without-internet-ain-worth.html>.
- [6] Wikipedia, http://en.wikipedia.org/wiki/E_pur_si_muove, accessed on Apr. 30, 2009.
- [7] E. Moglen, “Anarchism Triumphant: Free Software and the Death of Copyright,” *First Monday*, vol. 4, no. 8, Aug. 1999, http://emoglen.law.columbia.edu/my_pubs/nospeech.html.
- [8] Science Commons Protocol on Open Access Data, <http://sciencecommons.org/projects/publishing/open-access-data-protocol>.