



From Web 2.0 to the Global Database

TIMO HANNAY
Nature Publishing Group

ONE OF THE MOST ARTICULATE OF WEB COMMENTATORS, Clay Shirky, put it best. During his “Lessons from Napster” talk at the O’Reilly Peer-to-Peer Conference in 2001, he invited his audience to consider the infamous prediction of IBM’s creator, Thomas Watson, that the world market for computers would plateau at somewhere around five [1]. No doubt some of the people listening that day were themselves carrying more than that number of computers on their laps or their wrists and in their pockets or their bags. And that was even before considering all the other computers about them in the room—inside the projector, the sound system, the air conditioners, and so on. But only when the giggling subsided did he land his killer blow. “We now know that that number was wrong,” said Shirky. “He overestimated by four.” Cue waves of hilarity from the assembled throng.

Shirky’s point, of course, was that the defining characteristic of the Web age is not so much the ubiquity of computing devices (transformational though that is) but rather their interconnectedness. We are rapidly reaching a time when any device not connected to the Internet will hardly seem like a computer at all. The network, as they say, is the computer.

This fact—together with the related observation that the dominant computing platform of our time is not Unix or Windows or



Mac OS, but rather the Web itself—led Tim O’Reilly to develop a vision for what he once called an “Internet operating system” [2], which subsequently evolved into a meme now known around the world as “Web 2.0” [3].

Wrapped in that pithy (and now, unfortunately, overexploited) phrase are two important concepts. First, Web 2.0 acted as a reminder that, despite the dot-com crash of 2001, the Web was—and still is—changing the world in profound ways. Second, it incorporated a series of best-practice themes (or “design patterns and business models”) for maximizing and capturing this potential. These themes included:

- Network effects and “architectures of participation”
- The Long Tail
- Software as a service
- Peer-to-peer technologies
- Trust systems and emergent data
- Open APIs and mashups
- AJAX
- Tagging and folksonomies
- “Data as the new ‘Intel Inside’”

The first of these has widely become seen as the most significant. The Web is more powerful than the platforms that preceded it because it is an open network and lends itself particularly well to applications that enable collaboration. As a result, the most successful Web applications use the network on which they are built to produce their own network effects, sometimes creating apparently unstoppable momentum. This is how a whole new economy can arise in the form of eBay. And how tiny craigslist and Wikipedia can take on the might of mainstream media and reference publishing, and how Google can produce excellent search results by surreptitiously recruiting every creator of a Web link to its cause.

If the Web 2.0 vision emphasizes the global, collaborative nature of this new medium, how is it being put to use in perhaps the most global and collaborative of all human endeavors, scientific research? Perhaps ironically, especially given the origins of the Web at CERN [4], scientists have been relatively slow to embrace



approaches that fully exploit the Web, at least in their professional lives. Blogging, for example, has not taken off in the same way that it has among technologists, political pundits, economists, or even mathematicians. Furthermore, collaborative environments such as OpenWetWare¹ and Nature Network² have yet to achieve anything like mainstream status among researchers. Physicists long ago learned to share their findings with one another using the arXiv preprint server,³ but only because it replicated habits that they had previously pursued by post and then e-mail. Life and Earth scientists, in contrast, have been slower to adopt similar services, such as Nature Precedings.⁴

This is because the barriers to full-scale adoption are not only (or even mainly) technical, but also psychological and social. Old habits die hard, and incentive systems originally created to encourage information sharing through scientific journals can now have the perverse effect of discouraging similar activities by other routes.

Yet even if these new approaches are growing more slowly than some of us would wish, they are still growing. And though the timing of change is difficult to predict, the long-term trends in scientific research are unmistakable: greater specialization, more immediate and open information sharing, a reduction in the size of the “minimum publishable unit,” productivity measures that look beyond journal publication records, a blurring of the boundaries between journals and databases, and reinventions of the roles of publishers and editors. Most important of all—and arising from this gradual but inevitable embrace of information technology—we will see an increase in the rate at which new discoveries are made and put to use. Laboratories of the future will indeed hum to the tune of a genuinely new kind of computationally driven, interconnected, Web-enabled science.

Look, for example, at chemistry. That granddaddy of all collaborative sites, Wikipedia,⁵ now contains a great deal of high-quality scientific information, much of it provided by scientists themselves. This includes rich, well-organized, and interlinked information about many thousands of chemical compounds. Meanwhile, more specialized resources from both public and private initiatives—notably PubChem⁶ and ChemSpider⁷—are growing in content, contributions, and usage

¹ <http://openwetware.org>

² <http://network.nature.com>

³ www.arxiv.org

⁴ <http://precedings.nature.com>

⁵ <http://wikipedia.org>

⁶ <http://pubchem.ncbi.nlm.nih.gov>

⁷ www.chemspider.com



despite the fact that chemistry has historically been a rather proprietary domain. (Or perhaps in part because of it, but that is a different essay.)

And speaking of proprietary domains, consider drug discovery. InnoCentive,⁸ a company spun off from Eli Lilly, has blazed a trail with a model of open, Web-enabled innovation that involves organizations reaching outside their walls to solve research-related challenges. Several other pharmaceutical companies that I have spoken with in recent months have also begun to embrace similar approaches, not principally as acts of goodwill but in order to further their corporate aims, both scientific and commercial.

In industry and academia alike, one of the most important forces driving the adoption of technologically enabled collaboration is sheer necessity. Gone are the days when a lone researcher could make a meaningful contribution to, say, molecular biology without access to the data, skills, or analyses of others. As a result, over the last couple of decades many fields of research, especially in biology, have evolved from a “cottage industry” model (one small research team in a single location doing everything from collecting the data to writing the paper) into a more “industrial” one (large, distributed teams of specialists collaborating across time and space toward a common end).

In the process, they are gathering vast quantities of data, with each stage in the progression being accompanied by volume increases that are not linear but exponential. The sequencing of genes, for example, has long since given way to whole genomes, and now to entire species [5] and ecosystems [6]. Similarly, one-dimensional protein-sequence data has given way to three-dimensional protein structures, and more recently to high-dimensional protein interaction datasets.

This brings changes that are not just quantitative but also qualitative. Chris Anderson has been criticized for his *Wired* article claiming that the accumulation and analysis of such vast quantities of data spells the end of science as we know it [7], but he is surely correct in his milder (but still very significant) claim that there comes a point in this process when “more is different.” Just as an information retrieval algorithm like Google’s PageRank [8] required the Web to reach a certain scale before it could function at all, so new approaches to scientific discovery will be enabled by the sheer scale of the datasets we are accumulating.

But realizing this value will not be easy. Everyone concerned, not least researchers and publishers, will need to work hard to make the data more useful. This will

⁸ www.innocentive.com



involve a range of approaches, from the relatively formal, such as well-defined standard data formats and globally agreed identifiers and ontologies, to looser ones, like free-text tags [9] and HTML microformats [10]. These, alongside automated approaches such as text mining [11], will help to give each piece of information context with respect to all the others. It will also enable two hitherto largely separate domains—the textual, semi-structured world of journals and the numeric, highly structured world of databases—to come together into one integrated whole. As the information held in journals becomes more structured, as that held in many databases becomes more curated, and as these two domains establish richer mutual links, the distinction between them might one day become so fuzzy as to be meaningless.

Improved data structures and richer annotations will be achieved in large part by starting at the source: the laboratory. In certain projects and fields, we already see reagents, experiments, and datasets being organized and managed by sophisticated laboratory information systems. Increasingly, we will also see the researchers' notes move from paper to screen in the form of electronic laboratory notebooks, enabling them to better integrate with the rest of the information being generated. In areas of clinical significance, these will also link to biopsy and patient information. And so, from lab bench to research paper to clinic, from one finding to another, we will join the dots as we explore terra incognita, mapping out detailed relationships where before we had only a few crude lines on an otherwise blank chart.

Scientific knowledge—indeed, all of human knowledge—is fundamentally connected [12], and the associations are every bit as enlightening as the facts themselves. So even as the quantity of data astonishingly balloons before us, we must not overlook an even more significant development that demands our recognition and support: that the information itself is also becoming more interconnected. One link, tag, or ID at a time, the world's data are being joined together into a single seething mass that will give us not just one global computer, but also one global database. As befits this role, it will be vast, messy, inconsistent, and confusing. But it will also be of immeasurable value—and a lasting testament to our species and our age.

REFERENCES

- [1] C. Shirky, "Lessons from Napster," talk delivered at the O'Reilly Peer-to-Peer Conference, Feb. 15, 2001, www.openp2p.com/pub/a/p2p/2001/02/15/lessons.html.
- [2] T. O'Reilly, "Inventing the Future," 2002, www.oreillynet.com/pub/a/network/2002/04/09/future.html.



- [3] T. O'Reilly, "What Is Web 2.0," 2005, www.oreillyn.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.
- [4] T. Berners-Lee, *Weaving the Web*. San Francisco: HarperOne, 1999.
- [5] "International Consortium Announces the 1000 Genomes Project," www.genome.gov/26524516.
- [6] J. C. Venter et al., "Environmental genome shotgun sequencing of the Sargasso Sea," *Science*, vol. 304, pp. 66–74, 2004, doi:10.1126/science.1093857.
- [7] C. Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired*, June 2008, www.wired.com/science/discoveries/magazine/16-07/pb_theory.
- [8] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," 1998, <http://ilpubs.stanford.edu:8090/361>.
- [9] [http://en.wikipedia.org/wiki/Tag_\(metadata\)](http://en.wikipedia.org/wiki/Tag_(metadata))
- [10] <http://en.wikipedia.org/wiki/Microformat>
- [11] http://en.wikipedia.org/wiki/Text_mining
- [12] E. O. Wilson, *Consilience: The Unity of Knowledge*. New York: Knopf, 1998.