Robert Metcalfe, co-inventor of Ethernet and the founder of 3Com, is often attributed with the observation that the value of a telecommunications network is proportional to the square of the number of connected users of the system. This is known as Metcalfe’s Law, and it goes a long way towards explaining why we can create and realize so much value from the Web. As more users get online, the network gets more valuable, spurring more users to get online, and so on.

But we don’t have anything like the Web for data, especially scientific and scholarly data. Our personal data is mined by Facebook, by Twitter, by Google, to serve us with relevant advertisements to underpin many of the “free” services we access via smartphones and browsers. Yet those network effects - Metcalfe’s Law for data - remain elusive. We don’t have the users yet to achieve the moment in the Metcalfe curve where value generation breaks above the cost line - and the reality is, given the small number of scholars and scientists, if we depend on more people being trained, we might not ever make it. That’s why we need a Web of data that includes machines as users, not just people. Software has to be able to consume data in a meaningful way if we’re going to realize that value generation for science that we know from our social and consumer lives.

The Web of data would be a tremendously powerful tool for research. It would allow us to link together disparate information from unrelated disciplines, run powerful queries, and get precise answers to complex, data-driven questions. It’s an undoubtedly desirable extension of the way that the existing networks increase the value of documents and computers through connectivity.

However, making the Web of data turns out to be a deeply complex endeavor. Data - for our purposes, a catch-all word covering databases, datasets, and generally meaning here information that is gathered in the sciences as a result of either experimental work or environmental observation - requires a much more robust and complete set of standards to achieve the same "web" capabilities we take for granted in commerce and culture.

Unlike documents, the ultimate intended reader of most data is a machine - a piece of software that will process data and transform it into results that are comprehensible and meaningful to a human. Some classic examples include search engines, analytic software, visualization tools, database back ends, and more. There is simply too much data in production to place people on the front lines of analysis. When data generation per day scales into the petabytes, we just can’t keep up using the existing systems.

This machine-readability requirement is very different from the Web of documents, which was designed to standardize the way information is displayed by machines to people. Your browser doesn’t understand whether the web page you are reading is about physics or biology - but it knows precisely how to display the text on that page, whether it’s blinking (bad form!) or formatted using precise HTML (good form!). Machine readability for data means that the software running across data needs to be able to place that data into context, to know what other kinds of data it links to, by what other
software systems it might be usable, and so forth. And that information needs to travel alongside the data itself somehow, either by integration directly into the data or via a persistent URL that contains a stable description of the data.

Machine readability means we have to think, early and often, about the level of interoperability in any given chunk of data. "How "connectable" is it to other data?" should be the first question we ask of new data, because the level of effort required to make data connectable post-hoc is significant - frequently unbearable. Another way to think about this problem is as an interoperability issue: the connectability quotient creates significant pressures to build data that interoperates with other data ex ante, not post hoc.

Interoperability implies a level of rigor in the design of data that understands the intended use of that data is in a network context. Lightweight tagging and loose web 2.0 protocols that work for photographs don’t scale in a massively complex data world - it takes more rigorous and formal definitions, more persistent web names, more attention paid to writing things down as precisely as we can, because otherwise, the machines won’t know what to do. Making interoperability simple is in fact a deeply complex task.

There are three interlocking dimensions to interoperability in data: legal, technical, and semantic. By legal, we mean the contractual and intellectual property rights associated with the data; by technical, the standard systems (especially the computer languages) in which the data is published; and by semantic, the actual meaning of the data itself - what it describes, and how it relates to the broader world.

Each of these dimensions is complex on its own. Taken together, the three represent unsolvable complexity. The semantic layer alone requires an almost miraculous level of agreement on "what things mean," and anyone who has witnessed argument among scientists, be they economists of physicists, knows that even apparently simple topics turn contentious over matters as basic as definitions. Consensus on the technical layer is somewhat easier - the existence of the Web and the Semantic Web "stack" of standard technologies has begun to take a leadership position in data networking - but still difficult, long, and open to argument. One of the only opportunities for simplification is in the legal layer, where we can look to a broad set of successes in legal interoperability through the use of a simple, flat standard: the public domain.

The public domain exists as a counterweight to copyright in the creative space, but in some countries - especially the United States - as a first option for data that is not considered "creative." The public domain option currently underpins a wide variety of linked data that is already well on its way to achieving Web scale. From the International Virtual Observatory, whose members build an international data net on norms of "acknowledgment" rather than contracts of "attribution", to the world of genomics, where entire genomes and related data are harmonized nightly across multiple countries, the public domain creates complete interoperability at the legal layer of the data network, and serves as a foundation for the next layer of technical interoperability.
Interestingly we have yet to observe similar network effects emerging in cases where the underlying data is treated in a more conservative "intellectual property" context by using copyright licenses or database licenses inspired by copyright. Indeed, in the case of the international consortium mapping human genomic variation, the implementation of a “click through” license was found in practice to impede integration of that mapped variation with other public domain data, limiting the value of the map. The license was removed, the public domain option instated, and the database was immediately technically integrated with the rest of the international web of gene data.

The public domain option has been very successful in US Government approaches, and in private efforts whose goal is interoperability with the public domain like the Personal Genome Project and the Polar Information Commons. The realities of negotiation also require a set of “fallback” strategies when the public domain is out of reach - using an attribution-only copyright license along with attribution specifications that minimize unnecessary burdens on data reuse is one such approach, as this ensures that the complexity of the data rights issue is avoided, and the only risk is the “stacking” attribution in a future of massively parallel, linked data queries, which can be mitigated through good attribution policy in the license structure.

The legal element is of course just the beginning. The entities inside the databases themselves must be named and linked, in a standard way. Consensus on a dizzying array of technical standards must be achieved through working groups and hard won agreement. Semantic agreement - or disagreement - must be enabled where possible, and managed through savvy technology where not possible. It’s hard work.

However, it might be good to think back a bit to the emergence of the Web as a consumer force, and remember that the emergence of e-commerce and the digital commons wasn’t a foregone conclusion. It took the active participation of not just technologists, but entire industries that saw new markets in the connectivity of computers and documents.

Banks and credit card companies saw a benefit in moving transactions online, and supported the first blooming of electronic commerce. This foresight led to the development of secure web protocols, and digital site certificates, increasing the trustability and verification of online credit card orders. New business models emerged, bringing players like PayPal into the world, who in turn secured the millions of transactions moving across sites like eBay. This is the financial ecosystem that underpinned significant parts of Metcalfe’s Law for the Web.

And we need something similar for the web of data. The question is, what industries will step forward to lead?

I would propose that the scholarly publishing community - especially libraries and publishers - is in a unique position to lead. The publishing community has a deep and long-standing connection to the authors and the scientists who make the data, which is the first requirement to getting those authors and scientists to begin placing data on the
web. The publishing community has a tradition of peer review, of annotation and curation of research, which will be an essential component to making data useful on the web. And the publishing community has the ability to preserve information over the long term, by which we mean centuries, not years.

These three requirements - connection, curation, and preservation - must be satisfied if we’re going to see the sorts of increases in scientific and economic value for data that we see in other electronic networks. And there is no other community as well positioned as the publishing community to broker the necessary services and transitions.

Now, I am not proposing any unfunded mandates here. Rather I mean the opposite - there is a meaningful and sustainable business opportunity for libraries and publishers in the web of data. The US government alone is investing $150,000,000,000 in research. If we agree on the potential value created by making a web for data, it would only be a small investment as a percentage to pay for the services and tools needed to make data truly interoperate. And the recent move by the NSF to mandate data sharing plans gives us some clues about the future moves by public funders to prioritize data publication as part of a return on investment - there is going to be a market here. It’s simply a matter of finding business models that service it.

But one last word of caution, and one last word of hope. We cannot replicate the complexity and control of copyright-world at the data level. Publishers and scholars alike must avoid the temptation to treat data as property. If the entire system must begin with a complex set of legal terms and conditions, and be subject to the kinds of injunctions and property claims so familiar from the creative world, it is inherently unstable and unlikely to interoperate. We have seen the public domain option work, again and again, across the scientific disciplines. Implementing the public domain as the interoperability standard for the legal dimension of the web of data holds the greatest promise for scalability and long-term achievement of the network effect for data, as it permits the widest range of experimentation and development at the technical and semantic layers.