for example, but we should create several other such systems. And we just need to hope that some of these take off. It should not be very expensive to seed a large number of projects.

THE COMING REVOLUTION IN SCHOLARLY COMMUNICATION

I have reached the end of the first part of my talk: it was about the need for tools to help scientists capture their data, curate it, analyze it, and then visualize it. The second part of the talk is about scholarly communication. About three years ago, Congress passed a law that recommended that if you take NIH (National Institutes of Health) funding for your research, you should deposit your research reports with the National Library of Medicine (NLM) so that the full text of your papers should be in the public domain. Voluntary compliance with this law has been only 3 percent, so things are about to change. We are now likely to see all of the publicly funded science literature forced online by the funding agencies. There is currently a bill sponsored by Senators Cornyn and Lieberman that will make it compulsory for NIH grant recipients to put their research papers into the NLM PubMed Central repository.12 In the UK, the Wellcome Trust has implemented a similar mandate for recipients of its research funding and has created a mirror of the NLM PubMed Central repository.

But the Internet can do more than just make available the full text of research papers. In principle, it can unify all the scientific data with all the literature to create a world in which the data and the literature interoperate with each other [Figure 3 on the next page]. You can be reading a paper by someone and then go off and look at their original data. You can even redo their analysis. Or you can be looking at some data and then go off and find out all the literature about this data. Such a capability will increase the “information velocity” of the sciences and will improve the scientific productivity of researchers. And I believe that this would be a very good development!

Take the example of somebody who is working for the National Institutes of Health—which is the case being discussed here—who produces a report. Suppose he discovers something about disease X. You go to your doctor and you say, “Doc, I’m not feeling very well.” And he says, “Andy, we’re going to give you a bunch of tests.” And they give you a bunch of tests. He calls you the next day and says,

12 See Peter Suber’s Open Access newsletter for a summary of the current situation: www.earlham.edu/~peters/fos/newsletter/01-02-08.htm.
“There’s nothing wrong with you. Take two aspirins, and take some vacation.” You go back a year later and do the same thing. Three years later, he calls you up and says, “Andy, you have X! We figured it out!” You say, “What’s X?” He says, “I have no idea, it’s a rare disease, but there’s this guy in New York who knows all about it.” So you go to Google\(^{13}\) and type in all your symptoms. Page 1 of the results, up comes X. You click on it and it takes you to PubMed Central and to the abstract “All About X.” You click on that, and it takes you to the *New England Journal of Medicine*, which says, “Please give us $100 and we’ll let you read about X.” You look at it and see that the guy works for the National Institutes of Health. Your tax dollars at work. So Lieberman\(^{14}\) and others have said, “This sucks. Scientific information is now peer reviewed and put into the public domain—but only in the sense that anybody can read it if they’ll pay. What’s that about? We’ve already paid for it."

The scholarly publishers offer a service of organizing the peer review, printing the journal, and distributing the information to libraries. But the Internet is our distributor now and is more or less free. This is all linked to the thought process that society is going through about where intellectual property begins and ends. The scientific literature, and peer reviewed literature in particular, is probably one of the places where it ends. If you want to find out about X, you will probably be

---

\(^{13}\) Or, as Jim might have suggested today, Bing.

able to find out that peach pits are a great treatment for X. But this is not from the peer reviewed literature and is there just because there’s a guy out there who wants to sell peach pits to you to cure X. So the people who have been pioneering this movement towards open access are primarily the folks in healthcare because the good healthcare information is locked up and the bad healthcare information is on the Internet.

THE NEW DIGITAL LIBRARY

How does the new library work? Well, it’s free because it’s pretty easy to put a page or an article on the Internet. Each of you could afford to publish in PubMed Central. It would just cost you a few thousand dollars for the computer—but how much traffic you would have I don’t know! But curation is not cheap. Getting the stuff into the computer, getting it cross-indexed, all that sort of stuff, is costing the National Library of Medicine about $100 to curate each article that shows up. If it takes in a million articles a year, which is approximately what it expects to get, it’s going to be $100 million a year just to curate the stuff. This is why we need to automate the whole curation process.

What is now going on is that PubMed Central, which is the digital part of the National Library of Medicine, has made itself portable. There are versions of PubMed Central running in the UK, in Italy, in South Africa, in Japan, and in China. The one in the UK just came online last week. I guess you can appreciate, for example, that the French don’t want their National Library of Medicine to be in Bethesda, Maryland, or in English. And the English don’t want the text to be in American, so the UK version will probably use UK spellings for things in its Web interface. But fundamentally, you can stick a document in any of these archives and it will get replicated to all the other archives. It’s fairly cheap to run one of these archives, but the big challenges are how you do curation and peer review.

OVERLAY JOURNALS

Here’s how I think it might work. This is based on the concept of overlay journals. The idea is that you have data archives and you have literature archives. The articles get deposited in the literature archives, and the data goes into the data archives. Then there is a journal management system that somebody builds that allows us, as a group, to form a journal on X. We let people submit articles to our journal by depositing them in the archive. We do peer review on them and for the ones we like, we make a title page and say, “These are the articles we like” and put it into
the archive as well. Now, a search engine comes along and cranks up the page rank on all of those articles as being good because they are now referenced by this very significant front page. These articles, of course, can also point back to the data. Then there will be a collaboration system that comes along that allows people to annotate and comment on the journal articles. The comments are not stored in the peer reviewed archive but on the side because they have not been peer reviewed—though they might be moderated.

The National Library of Medicine is going to do all this for the biomedical community, but it’s not happening in other scientific communities. For you as members of the CSTB, the CS community could help make this happen by providing appropriate tools for the other scientific disciplines.

There is some software we have created at Microsoft Research called Conference Management Tool (CMT). We have run about 300 conferences with this, and the CMT service makes it trivial for you to create a conference. The tool supports the whole workflow of forming a program committee, publishing a Web site, accepting manuscripts, declaring conflicts of interest and recusing yourself, doing the reviews, deciding which papers to accept, forming the conference program, notifying the authors, doing the revisions, and so on. We are now working on providing a button to deposit the articles into arXiv.org or PubMed Central and pushing in the title page as well. This now allows us to capture workshops and conferences very easily. But it will also allow you to run an online journal. This mechanism would make it very easy to create overlay journals.

Somebody asked earlier if this would be hard on scholarly publishers. And the answer is yes. But isn’t this also going to be hard for the IEEE and the ACM? The answer is that the professional societies are terrified that if they don’t have any paper to send you, you won’t join them. I think that they are going to have to deal with this somehow because I think open access is going to happen. Looking around the room, I see that most of us are old and not Generation Xers. Most of us join these organizations because we just think it’s part of being a professional in that field. The trouble is that Generation Xers don’t join organizations.

**WHAT HAPPENS TO PEER REVIEW?**

This is not a question that has concerned you, but many people say, “Why do we need peer review at all? Why don’t we just have a wiki?” And I think the answer is that peer review is different. It’s very structured, it’s moderated, and there is a degree of confidentiality about what people say. The wiki is much more egalitarian.
I think wikis make good sense for collecting comments about the literature after the paper has been published. One needs some structure like CMT provides for the peer review process.

**PUBLISHING DATA**

I had better move on and go very quickly through publishing data. I’ve talked about publishing literature, but if the answer is 42, what are the units? You put some data in a file up on the Internet, but this brings us back to the problem of files. The important record to show your work in context is called the data provenance. How did you get the number 42?

Here is a thought experiment. You’ve done some science, and you want to publish it. How do you publish it so that others can read it and reproduce your results in a hundred years’ time? Mendel did this, and Darwin did this, but barely. We are now further behind than Mendel and Darwin in terms of techniques to do this. It’s a mess, and we’ve got to work on this problem.

**DATA, INFORMATION, AND KNOWLEDGE: ONTOLOGIES AND SEMANTICS**

We are trying to objectify knowledge. We can help with basic things like units, and what is a measurement, who took the measurement, and when the measurement was taken. These are generic things and apply to all fields. Here [at Microsoft Research] we do computer science. What do we mean by planet, star, and galaxy? That’s astronomy. What’s the gene? That’s biology. So what are the objects, what are the attributes, and what are the methods in the object-oriented sense on these objects? And note, parenthetically, that the Internet is really turning into an object-oriented system where people fetch objects. In the business world, they’re objectifying what a customer is, what an invoice is, and so on. In the sciences, for example, we need similarly to objectify what a gene is—which is what GenBank\(^{15}\) does.

And here we need a warning that to go further, you are going to bump into the O word for “ontology,” the S word for “schema,” and “controlled vocabularies.” That is to say, in going down this path, you’re going to start talking about semantics, which is to say, “What do things mean?” And of course everybody has a different opinion of what things mean, so the conversations can be endless.

The best example of all of this is Entrez,\(^{16}\) the Life Sciences Search Engine,


created by the National Center for Biotechnology Information for the NLM. Entrez allows searches across PubMed Central, which is the literature, but they also have phylogeny data, they have nucleotide sequences, they have protein sequences and their 3-D structures, and then they have GenBank. It is really a very impressive system. They have also built the PubChem database and a lot of other things. This is all an example of the data and the literature interoperating. You can be looking at an article, go to the gene data, follow the gene to the disease, go back to the literature, and so on. It is really quite stunning!

So in this world, we have traditionally had authors, publishers, curators, and consumers. In the new world, individual scientists now work in collaborations, and journals are turning into Web sites for data and other details of the experiments. Curators now look after large digital archives, and about the only thing the same is the individual scientist. It is really a pretty fundamental change in the way we do science.

One problem is that all projects end at a certain point and it is not clear what then happens to the data. There is data at all scales. There are anthropologists out collecting information and putting it into their notebooks. And then there are the particle physicists at the LHC. Most of the bytes are at the high end, but most of the datasets are at the low end. We are now beginning to see mashups where people take datasets from various places and glue them together to make a third dataset. So in the same sense that we need archives for journal publications, we need archives for the data.

So this is my last recommendation to the CSTB: foster digital data libraries. Frankly, the NSF Digital Library effort was all about metadata for libraries and not about actual digital libraries. We should build actual digital libraries both for data and for the literature.

SUMMARY

I wanted to point out that almost everything about science is changing because of the impact of information technology. Experimental, theoretical, and computational science are all being affected by the data deluge, and a fourth, “data-intensive” science paradigm is emerging. The goal is to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other. Lots of new tools are needed to make this happen.