Colorado State University Libraries
Conference Proceedings and Events
CI Days: Cyberinfrastructure 2010 in the Rockies
Transcription of The unreasonable effectiveness of open data, 2010

Collection: CI Days: Cyberinfrastructure 2010 in the Rockies
Title: The unreasonable effectiveness of open data
Date: 2010
File Name: CI_Days_2010_Wilbanks.mp4
Date Transcribed: November 2024
Transcription Platform: Konch AI

BEGIN TRANSCRIPTION

[00:02 - 01:22] Program Emcee: Okay. [indistinct chattering] Here we begin our afternoon program. And as Tom Peterson alluded to today, we've got two keynote presentations. And we did them in the order that would accommodate travel plans. And so we were doing the two one punch. And it's my pleasure to introduce John Wilbanks, who's the director of the Science Commons. And to introduce him, I'm going to start with the old parable. I I think it's from East India. I'm not quite sure. "He who knows not, he knows not that he knows not is a fool. Shun him. He who knows not, but knows that he knows not is a child. Teach him. He who knows but knows not that he knows is asleep. Awaken him. He who knows and knows that he knows is wise. Follow him." So we're going to have some wisdom imparted to us today by John Wilbanks. And I think he's going to awaken some things and some of us, and maybe disabuse me of participating in the first two of those categories personally. So please welcome John Wilbanks. Thank you, John, for joining us. [applause]

[01:28 - 02:19] John Wilbanks: That's a pressure filled introduction if I've ever gotten one. So thank you for having me here today. It's a pleasure to be here at CSU. And I'm going to try to cover a bunch of different topics here. But what I want to weave around is this idea of the unreasonable effectiveness of open systems. And I cribbed this title from a Google blog post on the unreasonable effectiveness of data. And when we think about open access, it's being really conflated a lot with open data now. And I want to try to [object clatters] untangle those two a little bit and see what they have in common and what they don't have in common. And then some ways that we might be able to provide infrastructure for openness, whether that's for literature or for data. Let's see if this works. [clicks] Okay. My clicker isn't working.

[02:21 - 02:21] Technical Support: Right now, if you pull it up and you pull it down, it will go.

[02:21 - 04:03] John Wilbanks: There we go. So in the open data space, I'm going to try to mainly focus on data because Tom was able to focus on the literature pretty well, is there's a lot of conversation about a data web or we need a web for data, or we need to make the web work for data the way that it works for documents. And this is a pretty popular meme right now, at least in the circles that I run in the open circles. And there's some reasons for that. Let's see. There we go. Okay. So Yelp is a good example of the sort of thing that people mean when they say a data web. So Yelp is a way of taking map data and sort of location data and mashing it up with yellow pages data and then letting users add reviews to it, which is a form of user generated data. So this is a form of the data web. And it's what happens when we make map data available on the web. Map data is pretty simple. It's based on latitude and longitude coordinates. We have a relatively common set of phrases for things like streets, and it's pretty easy to map things like street and drive over, but it lets us do this sort of mash up. Now here's another one. I just moved from, from Boston to the West Coast, and we didn't have this last winter, which is sort of annoying. So the mass Department of Transit is now making live GPS data feeds available from buses and trains, and especially the aboveground trains in a Boston winter. It would be nice to know when the next one's actually coming, so you can stay in the coffee shop until like, 30 seconds before it comes. So they opened up this real time GPS data, again tied to the latitude and longitude and the street maps of the city, and you get catch the bus as an application.

[04:04 - 05:54] John Wilbanks: So when people talk about a data web, this is what, in my experience, they mean. And this is data that is relatively simple public sector information. It's not data about complicated adaptive systems like our bodies or the world in terms of climate change or energy usage. And I think there's often a confusion between the things that the web will support, like these location-based, public sector information driven apps and the more complicated, complex adaptive system stuff, that is the sort of fundament of modern data driven science. And I think that's a problem that we need to surface. It's not a life threatening problem, but we need to be honest with ourselves about how hard it's going to be to do this sort of thing for biology or for climate change as we go forward. And so I know this... I heard this came up in an earlier talk, is this sharing of data for Alzheimer's disease. And it's important to note that this project started in 2003, a long time ago, about seven years ago. That's two grant cycles in bio world. It has stayed funded through those grant cycles, and it's done under a fairly complicated set of sharing agreements. If you actually go to the website, you can't download anything, right? That's sort of not evident in the article, but it's all governed by privacy and inspector access. Every access to the data set is logged and tracked. And so it's a good form of sharing, but it's a form of sharing that's sort of constrained. It's a walled garden. But what's fantastic about it is that, it's one of the first proof points that sharing data does anything

good in science, right? Some of Tom's comments about not wanting to share published articles that are already out there. We hear a lot of similar things about sharing data at all.

[05:54 - 07:34] John Wilbanks: And so this is one of the first proof points. And it really needs to be celebrated of the fact that sharing data in any way advances the scientific process more so than the sort of game theoretic. I'll share my data with you if you're worth sharing with you way. Because once you get your login, accept the terms of use, get accepted, then you can begin to do an enormous amount of work inside the system instead of having to regenerate all the MRIs and all the other sorts of brain data that were available. And so when you begin to think about a future where these various sorts of closets or walled garden sharing efforts emerge, how do we begin to link those into a network? And so when I think about a data web, that's what I'm thinking about. How do we take this and the project that's happening in Parkinson's and the projects happening in Huntington's, and stitch them all into a network? And the infrastructure for that is going to be a lot more complicated than the one that's required for Catch the Bus. And so the popular meme in the data web is to, in a talk like mine, I would lead you in a chant of raw data now and it's not Tim Berners-Lee did it for the first time, famously at Ted. And it's actually it's goosebumps when you watch it, but it's actually starting to happen at lots of other conferences. And again, my fear is that we confuse the simplicity and capacity we have for Catch the Bus and raw data now with what actually went into making that Alzheimer's data sharing effort useful. And so this is the sort of thing that the data what people will show you. And this is a map of all these different linked databases. And you've actually got like biological data and movie review databases in the same graph here.

[07:35 - 09:22] John Wilbanks: And they're on this... it's the same way that those are both on the web but it's not like there's a meaningful cross between the two and we can run traversal queries that get us knowledge from the connections between those. And that's what this is really about, is connecting data in order to get new knowledge out of what we've already done. And so I'm going to argue today at length to some of you, [wry laughs] that the infrastructure needs to be the focus going forward, and given that this is a cyber infrastructure day, I figured that was an appropriate theme to come with. But let's start with the assumption. So the assumption behind a data web is that if we connect enough data together, there will be a Metcalfe's law style increase in the economic value of that data. And so this is the sort of famous law that says, "As we connect devices and they comparably communicate, their economic value will increase over time." And so the real debate in open data is whether we are, you know, here or here. And I'm going to argue we're closer to here, at least when it comes to complicated scientific data. I think when it comes to consumer data, we're very close to here. And that's what you see in Catch the Bus. But having spent six years trying to integrate and liberate complicated biological data, and we've just now started doing this in energy

sciences and climate change, it's a lot harder than Catch the Bus, right? There's not this sort of agreement as to what things mean or how to name things. And there are very complicated regimes of privacy and economic concerns around that data. You just don't see in the sort of consumer map based applications. Not to mention that, with cell phones and automated GPS on everyone's cell phone, it's becoming easier and easier to generate open map data.

[09:23 - 11:09] John Wilbanks: So there's more and more free map data to build on that's foundational in a way that we don't... we just don't have that in the genome yet, for example. So that's my introduction. And my first point, I'm going to try to make four points today, is that if we're going to understand our way to data. We have to start by understanding a way to the scholarly literature because the metaphors for open access to data are coming out of open access to literature and open source software, even when they don't really fit. So I don't have to go into this town covered it beautifully. You guys know what open access is. Open access isn't a single thing. Open access, as we think of it, as a movement, as a success, is the product of multiple different actions that have been taken over time by a lot of different stakeholders. One of the most important classes of those is the funder mandates. This is the NIH public access policy. I could have shown you any number of funder mandates, the Welcome funder mandate, the huge funder mandate, and the smaller and smaller foundations that are beginning to implement similar ones. This is sort of leverage point one, the people that pay for the research want to have some rights over the knowledge products that they pay for. University mandates really fall into a similar category. Now, a second piece is the emergence of the Open Access journal as a business model competition aspect inside scholarly publishing. Now you don't see the people who are making tons and tons of money in scholarly publishing, radically shifting their business model immediately. This is being used as a leverage point for startups to gain leverage against the existing market, because the scholarly publishing market is incredibly stable. And in the absence of technical disruption and business model disruption, it's going to be hard for you to compete with nature or with Elsevier.

[11:10 - 12:54] John Wilbanks: But the advent of open access plus internet has created a business model competition. So it's not just the funder mandates, it's the emergence of these businesses, both non-profit and for profit, that compete with the established players using open access as part of the leverage. And then third and relevant to today is you have the emergence of the university thinking of itself as a player, in the dissemination of information, not just the creation of the archiving of that information. And so it's really these three things together that play into these unreasonably effective power of open access. If you think go back ten years to the year 2000 and you come in and you say 'There's going to be 6000 journals on the web that are completely free.' A lot of you might still say, 'Isn't the web where people have cat fan pages?' Right? It's really, you know, we didn't have

the web on our cell phones ten years ago, right? Websites were still... it was a debate whether you needed one or not, and how far it needed to go. Credit cards weren't really something you might want to use on the net. You certainly weren't going to bank on the web. That was just ten years ago. And in that time period, we've made this radical shift. And it's not because of any one thing. It's this confluence, right, of policy, new business challenges, technical capacities, and institutional support. And when you think about open data again, in this complex, adaptive scientific context, not the sort of end user consumer context, I don't think we have any of these in place yet. The way that we had them leveraged and ready to go for open access. We don't have the data policy, right? We don't have the business models for data, for annotating it, curating it, storing it, for getting it.

[12:55 - 14:41] John Wilbanks: We don't know how to store and forward giant data, right? I mean, we have tools going up into the night sky that are going to do petabytes every ten minutes, right. We have no idea what to do with that. Widespread sensor networks $100 genomes, you know. It's going to be go beyond supercomputing pretty quickly. And, you know, people might like, "Oh, it'll just sit in the cloud." And I say, "Well, to those of you who say that, remember that Google wave is already dead?" Right. We don't have a public institutional structure or technical system for storing giant, complicated data, much less the models that need to operate on that data so that we can turn it into something approaching knowledge and if we're lucky, wisdom, right? But I'll settle for knowledge. Now, this is the sort of thing that unreasonable effectiveness created. So BioMed central, which gives away all of their content for free under a license that doesn't make any restraints other than attribution to the author and to the journal was sold to Springer. They were bought. All the stuff was free, but they got bought. They were making about 15 million a year in revenues, US when they got bought. So this business model, I think, is one of the most important aspects of all of this. We had all this nice infrastructure for documents that the web gave us, and all it took was the really disruptive business model to help push this. Public library of science took advantage of that, and Darwin took advantage of that. The 6000 journals and the Directory of Open Access journals have taken advantage of that. And this is the license, right? So I work at Creative Commons. We approach the commons from the perspective of how do you need to construct it so that it works legally and technically. And this is the license that that BMC used. And I'll [thuds] go into this in detail in a little bit.

[14:43 - 16:20] John Wilbanks: And this is the point, I guess, of this first part, which is that literature we've had hundreds of years since Gutenberg to develop a culture and a society of documents, you know. And there were.. I'm going to quote Mark Lemley, who just came up with a great paper. I'm sure there were monastic Scribner's who were really upset that their business was going to be destroyed by the creation of books, and it probably was destroyed by the creation of books but it

wasn't bad for books as a whole. And the web in many ways came along and it disrupted the document business. But it was really good for documents, and it fit into all of the cultural norms of documents that we have. We read them, we look at them, we use desktops. We've got mental models and social models and scientific models and educational models for dealing with documents, even if we have lots of them and they're connected to each other. We don't have that for data. So let's quickly what do I mean by this in the scholarly context? For critical social norms, for documents, in the scientific and scholarship sense. So we need to know registration who had the idea and when. What was the date of submission? Certification is what peer review is all about. Dissemination was the traditional function of the publisher to make sure it got out to everyone's mailbox. And then the preservation of the scholarly record, not only through the library function, but just the whole idea of citation, that you can find that article that was referred to, even if it goes back 200 years, into a German language journal about chemistry. Right. Those norms are robust, stable, healthy in the document culture, right? Even if you are a scientist who's not really into peer review, right?

[16:21 - 18:03] John Wilbanks: This is, you know, we know how to do this for documents. You know, he's holding a document. He's not holding a data set. But we don't know how to do registration, certification, dissemination and preservation of data at this point. We have no idea what those four things mean in a world of really rich, sensor driven, robot driven data. Much less, again, the sort of computational models that turn those data systems into something that we can mentally access as people. And from a scientific perspective, we've had 350 years, give or take, since the first journal to do that. And so although the web has been disruptive and transformative, it's been disruptive and transformative in the sense of the format of the document, not in the sense of the underlying metaphor that we're dealing with. So, you know, that was 1665. This is 1874. This is a journal of Cambridge Entomological Club. Here is that same journal today. I could have shown you the Royal Society transactions, but the same norms apply. We've got a volume number. We've got a page number. We have a title, we have an author. And most importantly, we have the idea that you've been up all of your research into a paper, that the paper is the container by which we transmit the knowledge around. And how does giant data fit into that metaphor? Right. So oh, fits into that [paper rustles] digital version. It's one of the core access, sorry, core benefits that Oh had is that it fit into the way that we already thought about things. It just sort of righted the ship with the way we thought that the internet actually already worked. Now these are... this was the Elsevier paper of the future. It looks an awful lot to me like a paper of the past, but with paper from the past, but with video, right, and audio.

[18:03 - 19:46] John Wilbanks: See, there's an author interview, but otherwise it's the same sort of paper, right? It's this introduction, results, discussion, experimental procedures, figures, references,

authors comments, acknowledgements. That looks like a paper of the past that's been formatted into HTML effectively. But the idea is that the knowledge is still contained in the paper. And here's another one, which is a semantically marked up paper. I like this more. I'm a semantic web geek. But again, the container and the metaphor is still that of a paper, not of some sort of complicated object which has data associated with it, which has commons associated with it, which might come from 4 or 5 different sources that need provenance and tracking and citation, each on their own. It's like this idea that we take a polaroid snap of the of the knowledge, and that's our paper. So that's my little reality check about big data and science, is that, the metaphors that containers and the norms that we've got for big data and science aren't really ready for the sort of flood of data that's coming. And the flood is coming, because our capacity to measure is far outstripping our capacity to use and reward. And that's not really going to stop. And it's important to remember, I graduated from college in 94. If I were the sort of person who got a PhD and I had gotten it quickly, I'd have gotten it in 99, I finished my postdoc in 2003. If I was lucky, I get my first grant, survived through to 2007. Get another grant, which I'm wrapping up now. So I'd be about ready if I was doing well for application as an assistant faculty member in a tenure track position. That entire transformation has taken place in the span of my single early career, right?

[19:46 - 21:25] John Wilbanks: No one talks about this, but it's one of the biggest resistance points to this sort of change, right? I'm right on the cusp of getting tenure or tenure track. Am I going to stop and annotate my data for someone when I'm never going to get rewarded for it? Am I going to stop and curate my data? Am I going to go to the effort to fight with the university and the IRB to get approval to release information? Right? That sort of infrastructure that makes it easy to to have an argument about whether or not the document goes into the institutional repository. It's almost too early to even have that about data in some ways. Now data has three sorts of things that are going to have to emerge, right, in addition to those registration, certification, dissemination and preservation facilities. We're going to need to have integration as a new capacity that we don't have right now. So this is the International Polar Year. Each of these honeycomb things. I'll zoom in. These are each projects just, you know, monitoring human ranger for migrations. There's 408 projects in there. They all generate data. Greening of the Arctic 139, hydrological cycle 104. I would wager dollars to doughnuts that none of these people are collaborating with each other to make sure their data in or operates. So post-hoc integration of data is going to need to be a primary facility that we provide, either through the business models or the institutional support. If we're going to take advantage and make a data web actually happen. Annotation, like this one is enormous. There's a great apocryphal story of astronomical observations sort of gathering dust in D.C. because no one wrote down what part of the sky they photographed.

[21:26 - 23:00] John Wilbanks: And you hear this again and again and again across the sciences that without proper annotation, the half life of a data set is very short. The useful half life of a data set to someone else is really quite short. As I'm non-scientist and a coffee junkie, I think of it in simply in the context of coffee. If we're going to meaningfully make this sort of data annotation possible, we have to start with a very basic idea, which is we have to get the names for things harmonized. And this is again how far we have to go with the infrastructure. These are all names for black coffee right up here, at least at the top. Right. That's a kind of black coffee, coffee, coffee, coffee, coffee. And then down here we've got, you know, mocha. So it's got chocolate latte, that's got milk, that's got water in it. But these are all ways that different cultures in different places, at different times talk about the same thing. Now, if you scale this out to the genome, some of the genes have as many as 100 synonyms, and they exist in as many as a thousand databases. So before we can do meaningful annotation about the meaning and the knowledge that's inside these things, we've got to go in and do a basic gardening effort to get the names harmonized, linked and shared, because that begins to give us the ability to do this. It's one of the major, I think, service opportunities out there as the data web explodes. If I was going to start a business, I'd try to figure out how I could automate annotation in a way that was moderately accurate, because I think you could make a lot of money. And then Federation is the last big one. The data is going to be too big to push across the wires, especially as people move more and more to wireless.

[23:00 - 25:05] John Wilbanks: So we're going to need ways that work to actually query data in 100 different places at once, as if they were all in the same place, much the way we Google the web. But the systems right now don't support that very well. So bringing all of this together is something that the web is going to have to support over time, not in the individual repository, but the repositories, the institutions, the scientists and the companies are going to have to use common interfaces so that we can actually begin to run federated queries across databases that weren't designed to work together. These are the three services we don't have that we're going to need to have before we can really have a meaningful open data movement, I think. So that's part one, and I apologize for being a bummer about it, but I actually am very optimistic about all of this stuff. I just want to make sure we get it straight, more honest with ourselves. So one is that not all data is created equal. So this is one, in my opinion, one of the shining examples of meaningful, hard open science data, which is the International Virtual Observatory Alliance. Now you can go to their website. Everything's in the public domain. You can download it. Everything's just run by norms and they have like a requested citation form on it. And these are all the technical specifications they had to write and all of their versions to harmonize the data from the observatories around the world. If you've ever been involved in a single technical specification effort, you know how this represents a world of pain, right? This was uncountable hours of conference calls and arguing over minutia and personality conflicts, that almost

are unthinkable that they got this done, right? It's incredible. And it works, right? Here's a single one of their technical notes, which is an ontology of astronomical object types. I read this and I got, you know, chills thinking about the arguments people had over how to define what a quasar was, right? Because, you know, if you have five astronomers in the room, that there'll be three opinions about how to do that in a technical way.

[25:06 - 26:51] John Wilbanks: And this is what I mean when I talk about the infrastructure. This is not abstracted outside of astronomy, but I think it's indicative of the scope of the work that's needed to meaningfully connect these sorts of big, heavy science databases and run federated queries. Here's their standards process they had to come up with to write all of these technical standards, right. You've got endorsement revision. It starts with a note, a working draft, a working group, a recommendation, an executive review, a recommendation, a new working group and then it's a standard, right. And this is the submission log starting in 08, 2009 through to June 2010, Right. I think we're a pretty long way away from this in most fields, right. But this is what it really takes. And so this is data that's actually right now more than equal. This is data that's not only usable., it's annotated. There's a process. It's curated. You can federated query it. It's all been put together for you, right. But it's not raw data, right? It's not raw data now. And it's actually, this doesn't roll off the tongue nearly as well, right? But this is what they actually did, right? Metadata and standards and consensus and document submission and archives now. And this is the sort of thing that you're going to see coming out of places like the national labs in the United States. You're going to see it coming out of certain NSF grants, like the data net grants that they've been very slowly, slowly promulgating, is exactly this sort of raw data now approach, which is much slower, but in the end much more powerful because it allows us to begin stitching together data from a lot of different complicated processes and asking interesting questions of it not when's the bus coming?

[26:52 - 28:48] John Wilbanks: But, you know, where should I put my desalination plant, given that the increase in global temperatures appears likely to lower snow melt in the Pacific Northwest enough that we can't have new urban settlements near Seattle as of 2050? Right? That's the sort of complicated climate modeling question that can only happen with this sort of statement. And that's because all the data that we need to answer that question was gathered by different people at different times, [recording rattles] for different reasons, under different grant regimes. And it's only through these sort of large scale [splats] institutional efforts that we can really stitch it all together. But if we do it right, we don't have to repeat it each time. We do it once for any given data source, and we move on. And then they begin to connect because we're using the same names for things, we're using the same data formats for things. And because ideally there begins to be an institutional commitment, a business model incentive, and a policy mandate to do so. So I'm going to shift gears

a little bit and talk about once we have that data, what regulates it? Because it's important to think about data in the sense of how it gets regulated. And it's not just intellectual property for data. There's privacy, which is a totally different regime from IP for data, as the technology for data, and then there's the policy that really governs it. So IP is what people expect me to talk about since I come from Creative Commons. You know, the reality is, copyright in relation to data actually doesn't have a very big role. In the United States at least, raw data is in the public domain because it's not a creative work. IP can interfere with these other aspects if you try to license it the way we license IP in documents or in software. If you try to make copyleft style arrangements happen in data, you can actually wind up crossing over with privacy because the privacy says you can't release this data. And then the copyleft license says you're required to release this data since you made a derivative work.

[28:49 - 30:24] John Wilbanks: And you wind up in a in a zone where the dataset is unusable because the conflict between privacy and intellectual property rules. Technology we sort of talked about a little bit. We did this in science commons, we tried to take about a 100 data sources on [unintelligible] and life sciences, and actually wire them all together. Unlike the [unintelligible] way, we did not have enough bandwidth to do consensus. So our our rule was to give it all away and if you disagree with us, you can fork the code. But it took us about 3 years to wire together just this minimal set of data. And now that it's done, the 'n' plus one is pretty insignificant, speaking that of adding new data sources in a week, if we need to and anyone else can add it in a week if they need to. But from a technical perspective, the standards and the infrastructure are weak enough [unintelligible] had to do the work and [unintelligible] took years, not weeks or months. So to me the appropriate metaphor is the internet, and not the web. I think when it comes to data, we are much more in a world like the NSF Net or the [unintelligible] than it is that we are in the world, where we ever just gonna turn the corner in two years and the webflow data is going to explode. I think we at least need to have a browser for data that's meaningful, widespread and popular, before we get that explosion. Because Chrome, Internet Explorer don't cut it for data. Policy- we have talked about this. We have this great mandate in the NIH for open access to literature and elsewhere. Here's their equivalent for data. Basically it says please submit a data sharing plan. But there's no guidelines as to what a data sharing plan needs to be.

[30:24 - 32:12] John Wilbanks: I assume the data sharing plan could be I plan to not share my data. [audience laughs] That's a plan. It's not a satisfactory one, but it's a plan. There's no metrics by which the study session has to evaluate that data sharing plan. They don't publish the data sharing plan next to the grant ID, so we know what that investigator promised. We can't track whether or not the investigator is complying with the data sharing plan he or she submitted. There's no tracking, follow up, accountability or transparency. And this sort of is the case for almost every data sharing

policy that's out there and that makes them effectively toothless. It makes them opt in. And Towne showed us exactly how effective opt in strategies are. And that leaves privacy. Privacy is so complicated that it's almost not worth going into, except to note that privacy varies radically across nations, and it varies radically across classes of data. And privacy is probably the single most complex and difficult thing to solve in an open data context, because in a world where there is enough data, almost everyone is identifiable. I was at a computer freedom and privacy event a couple of weeks ago in the Bay area, and they had the guy who broke the Netflix algorithm next to me. So Netflix gave away all these movie reviews and said, you know, "Improve our recommendation algorithm and we'll give you $1 million if you can beat what we've got right now by 10%." They de-identified all the data, but these other hackers came in and said, you know, "By cross-referencing the de-identified data you gave us with public information from other movie review databases, we can uniquely identify essentially everyone in the data set." The quote that stuck with me is if I know the origin of five pieces of your clothing, I can identify you, if you use credit cards to pay for them, right?

[32:12 - 33:38] John Wilbanks: That is a major, major issue and no one is really ready to deal with it yet except the and the people who are out there ahead of it are the people who don't want there to be any privacy like the folks who run Facebook. Right? That's part of their business model, is to have access to this really fine grained information about you and be able to monetize that. And right now, that's the dominant business model around open data is advertising. So when we ask, what's a regime that brings these things together at web scale? I don't think it's going to be something that we've seen before. I think it's going to be something that is similar to what we've seen in open source software, open access, but uniquely mapped to the data world. And so in the IP context, one of the proposals is let's do a very complicated copyleft regime. This is actually from the lawyer who drafted the sort of primary copyleft license for data. This is the simplest graph I can find on it. And that's why we decided Creative Commons not to do something like this. We actually decided that to make it simple, the best thing we can do is actually just make the IP go away. And to focus on this concept of the commons as a way to create an operating system for open access to data, because that's what's worked in software and in literature. And if we localize it, import it for data, we think it can work. So the commons is not a tragedy, although it's frequently associated with this concept of tragedy. A digital commons is a place where the resources are not rivalrous where if I have a song, you can also have the song.

[33:38 - 35:09] John Wilbanks: A traditional commons would be one where we've got the grass and I have a goat, and you have a goat. But if each of us keeps adding more goats, all the grass gets overgrazed. And that was the idea of a tragedy, is that none of us had an individual incentive to

protect that common space. Each of us had an incentive to destroy it, actually. But that doesn't really work in the digital context, and there's more and more evidence that in a digital sense, the commons is actually a comedy. It's something where you get unexpected wonderful things instead of unexpected horrible things. But all of the rules we have around the law are set up in the context that it's a tragedy, and that's where Creative Commons really comes into play, is we're trying to create these public footpaths across the property rights, where it's easy to begin building complex systems, copyrighted systems of content, data systems, materials, transfer of stem cells, and so forth. Now it's private, it's voluntary and it's pre negotiated. Those are key elements of a Commons. It's technically enabled, and we actually put it into the web itself. It shouldn't be something you have to do separated from the internet. This is our home page if you feel like going to visit us. We're a nonprofit organization with operations in over 70 countries. Everything we do is free. Everything I talk about is free to use, free of charge, and purely voluntary. And what we are most known for is our copyright licenses. You saw this earlier for BioMed central. This is a license that essentially instantiates the open access ethos. You are free to copy, distribute and transmit the work. You're free to adapt it as long as you give attribution. It's available in a lawyer readable form as well. [audience laughs]

[35:12 - 36:48] John Wilbanks: And it's available in a machine readable form, which turns out to be really important. So this is the HTML metadata that allows Google, Yahoo, Microsoft Office, and lots of other software systems to search for things based on the rights associated with them and not just the content of those things. And this is going to be important, and I'll come back to it in a minute. We've also expanded the commons into things like biological materials. This is a commons agreement for things like stem cells and plasmids. The idea is that we would like to standardize all sorts of regimes where currently this unnecessary negotiation taking place for fairly basic uses of tools. And so since stem cells are something where I can grow lots of them, like in a greenhouse, we actually went out and crafted these agreements, and they're sitting on about 10 lines of stem cells. You can order right now for $85 and go ahead and make commercial uses, but the project's actually going to be at 100,000 in the next five years. So there'll be 100,000 sets of stem cells that as long as you're a real researcher, you can essentially one click, order them and start doing your research right away instead of begging for the materials. We've just recently branched into the patent licensing space. This is actually up for public comment right now. This is our model patent license project, which is attempting to again, bring the concepts and the principles of the commons to a place where they haven't existed before, which is the rights to actually prevent people from making and using and selling technologies. Because if you don't go all the way to this, you can you can unleash a lot of research, but it's harder to actually help individuals in the real world get technologies and products into their into their lives that make things better. [clicks]

[36:49 - 38:31] John Wilbanks: And then for data, as I alluded to, we've created a tool called CC0 that eliminates the IP from those regulatory factors on data. And the whole idea is that the best thing we can do, given how complicated the privacy, the technology and the policy spaces are, is eliminate the complexity of the law entirely through a public domain approach. And then these are the jurisdictions that we've ported to internationally. The idea is that the commons has to be an international regime so the copyright licenses and many of our other tools actually get translated legally and linguistically into these regimes. So you can upload a song in Brazil, you can download it in the United States and Inter operates from a legal perspective. And we've had some great success on this. We don't have the numbers. We actually changed the way we started counting. So I'm not including our 09 numbers yet. But the last time we did a formal count, we were over 500 million objects on the web under our licenses, and we expect that has gone up significantly over that. There's 130 million photographs on the Flickr website alone at this point. And this is, you know, for a small NGO, this has been a pretty remarkable growth curve. Wikipedia, we just go through... I'm going to run through a few of our greatest hits of adopters. Wikipedia is a pretty big one. The white House actually changed to Creative Commons licenses during the inauguration speech in 2009, which was pretty cool. There's more than a thousand journals worldwide under CC licenses. We believe that number to be somewhere around 1500 right now. The personal Genome Project. This is where the stem cells are. These are the major adopters of the materials transfer agreements along with... there's a major repository of neurodegenerative physical research tools for Huntington's disease as well.

[38:31 - 40:05] John Wilbanks: This is actually pretty neat because it's the 100,000 lines of stem cells, but also for those individuals, they're full sequence genomes and health interviews. So you'll be able to say in five years or so, "Get me all the stem cell lines of Caucasian males in their late 30s that eat too much bacon, and test a cholesterol drug out in their stem cell lines before you go into the clinic." Right now, that's something that's really prohibitive to do but the commons enables that much the same way it enables Wikipedia content to be shared. And these are the users that we've had of our patent licenses so far, with Nike being the biggest one. Again, companies are starting to recognize on the patent side, the value of opening up their portfolio in a standardized way, in many ways better than the university community is. It's been pretty remarkable. But the idea is that the Commons cuts across all of these things, right? It's a way of looking at the world of property and the web and saying, we can bring these two things together through standardization and good technology and radically increase the usage of the knowledge, the data, the knowledge products that encode a lot of that knowledge, like patents and biological materials, because that's one of the best ways to actually accelerate the creation of new knowledge, and ideally, the use of that knowledge in a meaningful way. If we can eliminate those artificial barriers and increase the usage, that's one of

the best non miraculous ways we've got to increase the throughput of our scientific and innovation systems. And that's about making that law go away on data coming back to the theme of the day. Now, we've been working on this for six years. Our first protocol came out in zero seven on it.

[40:06 - 41:53] John Wilbanks: This is something you should know about, which is called the Panton Principles. So the folks who actually wrote that license for copyleft on data, the Open Knowledge Foundation and Creative Commons and others came together to agree that actually, even though they had written that license for data, it wasn't something that should be used in the sciences for data, which is that if science is publicly funded and referred to in a published article, it should be put into the public domain and made available without restrictions other than those required by privacy. It's a pretty big sort of consensus statement in the open data space. Nature has come out in favor of the approach we've laid out, including CC zero. GSK, the pharmaceutical company, not a bunch of hippies, became the biggest corporate user of CC zero a couple of months ago when they deposited essentially 13,500 chemical structures that are bioactive against the malaria genome in the public domain, using CC0 as the tool for that. Sage Bio Networks. This represents about $200 million worth of [chair scrapes] Merck's disease biology data moved out into a non-profit organization made public again under the public domain tool. The polar information commons that honeycomb I showed you, the output of that, again available from an international affiliation under public domain rules and norms. The Tropical Disease Initiative. This is essentially a set of data and information that helps investigate rare tropical diseases. Again, there's this momentum behind the idea that public science and the public domain for data is the best way to go from a legal perspective, because of the understanding that even after we get rid of the IP and the law, we've still got the heavy lifting to do, which is the technical aspects of integration and the social aspects of actually deciding what to do with the data.

[41:54 - 43:28] John Wilbanks: There's even a file sharing network for genomes now out of the University of Michigan. It's like BitTorrent, except for genomes, and it begins to get at one of those federation issues. You can actually, you know, have a BitTorrent network of petabytes of genome data that allows you to move that information around the network a little more efficiently than making just, you know, CSU libraries bear the load. And this is the last one, is that even in the European Union, where they love their database rights, we've seen some preliminary use of the tool at the European Molecular Biology Lab. So I'm beginning to get towards the end. That's what the Commons really is. And I was really talking about the law, but I had alluded to the idea that we knit this into the web itself. And so what's interesting about a digital commons is that it allows us to actually integrate these rights and capabilities into the fabric of the web in a way that we don't have to think about them from the perspective of the user, and that's really important. So I just got back

from a long trip to Indochina. And I can tell you that one of the best metaphors for non interoperable technology is the electric socket. [audience laughs] All right. So here's here's just three. And so when I talk about the technical aspects of this, what I really mean is interoperability. And I'm an evangelist, unashamed of the semantic web. And it's not because I think it's the best thing out there if we design from scratch an ideal world. Right? Just like I wouldn't say that this is the greatest thing on earth. We could just. We should just start with all using the same plugs. But this allows us to deal with the world that we have.

[43:29 - 45:09] John Wilbanks: And from a technology perspective, that's what a digital commons can bring. It's not just the legal aspects, but it's a commitment to standardizing the way we do technologies and names, ontologies, data formats and everything. Because that's what really lets us realize the power that the legal tools give us. Because if we don't use the technical interoperability, then the legal interoperability sort of creates an opportunity that's unrealized. We practice what we preach. We've written an actual language you call the CC rights expression language, which expresses all of the obligations that are in all of our licenses, whether patent licenses, data tools, materials transfer agreements, copyright agreements, and a machine readable standard semantic web format. When we put out deeds for things like data, this is the new CC0 deed. There is no... wrong button. There's no actually cut and paste citation. The idea being if you're a user, we don't know how to site data yet. We don't have journals for data. Well, we can actually make this essentially a URL based cut and paste where the user doesn't even have to know. I go, I see the data set, I click, I get this deed, it tells me what the rights and the regulations are, and I can cut and paste the URL to know that I've actually fulfilled my obligations. That sort of technical capacity makes it more likely that someone starts building metrics on the reuse of data, which tilts towards the policy and the business models and the institutional support, because right now we have no idea of how to track that. But we've got a nit citation into the web because right now it's knit into journals. And if any of those elements that we talked about earlier are going to come to pass, they've got to be part of the web, not just part of journals.

[45:10 - 46:49] John Wilbanks: And we've done a lot of boring standards work, so we've spent a lot of time on World Wide Web Consortium standards groups, the Web Ontology Language to report the Technical Architecture group. And we've created what we call a shared names project to create very boring, standardized URLs for entities inside databases. And all of this stuff is available on our website if you're interested. And I alluded to this earlier, we've sort of eat our own dog food, and we've actually gone ahead and used all of these tools to promulgate shared, integrated, federated data resources. And so the the point of all of this is that we're going to need a long term infrastructure commitment to this for data. Creative Commons is a non-profit operating in a very

uncertain economic environment. It's almost tragic that we're an infrastructure provider. That the government hasn't seen fit, that universities haven't seen fit, for the most part, to make this a priority. And so the hope is that through the advent of the digital commons, by making the legal rights available, by making the technology to track and value some of these things available, that the policies and the business models and the institutional support will begin to emerge. But when I think about it, I think about it, not again at the scale of the web. I think about it when I think about the Tennessee Valley Authority or when I think about the Eisenhower Highway System commitment, when I think about the Arpanet That's the level at which I think we need to contemplate infrastructure for data. If we're going to build a society that's actually data driven instead of one that's sort of focusing on iPhone apps. And that's just to build it, right? The discovery comes after let's go back to that Alzheimer's example. They started sharing three, seven years ago.

[46:49 - 48:23] John Wilbanks: And the first really big, you know, New York Times where the discovery is now. And so we cannot have an impatient attitude which the web has trained us to have, which is we live in a 24 hour news cycle world. We want instant returns on investment. Infrastructure takes a long time to pay off, and it's often in ways that we didn't expect. That's usually, in fact, the point of infrastructure. Okay, so I'm almost done. I've used up more than my amount of time, I think. Don't take this as a buzzkill, right? This is not bad, right. It's really good that we're actually at a point where we can talk about these things, because we're beginning to understand the problem. And that's the first step towards solving it. We didn't even begin to understand this problem even 5 or 6 years ago. But now that we're beginning to live with a lot of this data and live in a world of big data, we're starting to learn some of the lessons and get some of the use cases we need to actually solve these problems. So I'm actually enormously optimistic about linked open data and the data web, no matter what it might seem. We always overestimate the value of technology in a short term, and it's always disappointing, [woman coughs] but we almost always underestimate its impact in the long term. And so if we continue to make these investments in the infrastructure and in the data, we're going to be stunned by what happens. It's not going to be what we thought. Even us futurists that are sort of paid to think about these things. I hope I'm wrong because I doubt I can imagine the coolest thing that can be done. All right. That's the good news in all of this. And so I go back to to Metcalf's law.

[48:24 - 49:20] John Wilbanks: You know, it doesn't really matter whether we're here or here on the curve, as long as we're on the path. And as long as we have the commitment and the will to stay on that path. But if we can push for the open access regime policy, institutional commitment and business models and work day to day on things that enable one or more of those three things to come to pass, then we're going to keep moving up the curve. And so in a world in which we're going

to have a Metcalfe's law style path for data, the opposite of open is actually not closed. And this goes back to the unreasonable effectiveness, right? In a networked world, the opposite of closed of open is broken because if it's not accessible, if it's not discoverable, if it's not usable, it's going to be ignored. It's going to essentially be broken. And we're not there yet. But that's certainly where we're going in the open access movement is one of the best guide stars we've got in that. And with that, I'll stop, say thanks and take any questions you've got. [applause]

[49:29 - 49:45] Program Emcee: Thank you very much, John. Questions? [wry laughs] You're all stunned, right? TMI? Too much information?

[49:48 - 49:52] John Wilbanks: Oh, come on. There's got to be one. [microphone tilts] No?

[49:58 - 50:19] Woman 1: I guess after hearing this, I'm wondering if, you know, we're thinking about putting data in repositories, and I'm wondering if that's something that should wait for a while or should it... is it something that needs to have something else happen first? Or what's your thought on that?

[50:19 - 51:33] John Wilbanks: I mean, I would argue for starting now. I mean, I think that that's that's one of those actions that supports the institutional support lag of open. And it's a place where there's also not a business model competitor. You don't have the journal industry fighting you on data. And so I think that, that's actually green fields for institutional repositories. It's a place where I think there's enormous value to be added, but it's going to require a pretty significant interaction with the faculty member that's generating that data so that it's annotated, curated and preserved properly. But I think that's, you know, that's one of those places where there is no infrastructure for it. And that's a great place for a library community to step up. It's a place where you can really clearly demonstrate the value of a digital library in an institutional repository facility, as opposed to sort of contrasting it to what we're going to have, you know, this mishmash of the articles that our faculty wrote and everyone's going to hate us and fight us. If that becomes a service you can provide to faculty members, I think that increases the engagement and makes it easier to get that institutional policy over the long term. So I would... I hope I didn't indicate that you shouldn't start now. I think you absolutely should start now to the extent that you can.

[51:37 - 51:00] Man 1: Yeah. Just really quickly, which. I'm sorry. I missed part of your talk. You might have said this really clearly, but what do you think is the the big difference, if you had to summarize it [object rattling] between saying, "I'm going to release this object where object might be code or technical information or whatever into the public domain versus releasing it under a Creative Commons license." Can you explain that really quickly?

[51:00 - 54:09] John Wilbanks: Sure. So the public domain is an absence of rights viewed one way, the other way it's that all rights are granted. So the only prohibition on using something in the public domain would be ethical or professional plagiarism, right? Plagiarism isn't a legal crime. It's a professional transgression. So releasing something into the public domain means waiving all rights or granting all rights. Now, data in the United States is naturally in the public domain already. So even if you put a Creative Commons license on it, it wouldn't affect anything because I'd still be able to take it right. It's in the context of databases, have you ever played operation when you were a kid and you could get the bone out without touching the metal? You can extract the data from the database without touching the sides of the database. It's in the public domain, and you can do whatever you want with it. Now, using a Creative Commons license or an open source software license is what we call a some rights reserved approach, where you grant free... in a Creative Commons sense, you grant the right to make and distribute copies of the object, but you can retain the rights of attribution, the right to prohibit commercial use, the right to prohibit derivative works, or the right to require that a derivative work be made available under the same terms that it was made available to you. So that's the difference. It's the difference between sort of all rights granted slash no rights reserved, and a regime in which there are some rights reserved. There's a conditional grant. The other big difference is that under a Creative Commons license, if I make an article available to you under the condition that you give me attribution, and then you fail to give me attribution, you are in violation of the copyright, and I can sue you for infringement, whereas if it's in the public domain again, if you do that, I can simply call you on it in a professional context. But I can't call you at court. So that's the big difference. And we argue for a public domain in the data context. We think that the main issue with the public domain and the document context is that it takes too long for things to enter into it, but not that there should be sort of, you know, we're not advocating that open access journal to start using a public domain tool.

[54:14 - 54:26] Man 2: Faculty having to wait seven years to get some of the benefits from some of the extraordinary effort they're going to have to put in up front. Is that's something that's going to need to have to be funded by the funding agencies?

[54:27 - 55:20] John Wilbanks: I think so. I mean, if you look at the, again, that Alzheimer's story from today, it was well funded from the beginning and it had this very visionary set of leaders. I mean, the one I know is Doctor Weiner at UCSF. I mean, they really believed from the beginning that this was the way to advance science. They were unreasonable people and demanded that it be done. And they were persuasive enough scientists to get large amounts of money from the NIH to do it. You know, it really is a lot and right now about people deciding this is the way to do it and making a commitment and staking a piece of their careers on it. And that's tough to do when you're either on

the cusp of your career or comfortable in it. And, you know, the good news is that from the Public Library of Science to the ADNI that we saw today to many other efforts, people are beginning to be unreasonable about open systems. And that's one of the things that gives me great hope.

[55:28 - 55:00] Woman 2: [inaudible] Thinking about what the people in this room can do and what we can do on our campuses. It strikes me that one of the most concrete things we as a library community, can do is take a lead in establishing the use or requirements for DOIs for our for our data sets. If you can't find a data set, if the data set isn't going to be around, none of this is going to happen. And I don't know if that's... I'm just trying to think about, in the aftermath of this fire hydrant approach that you-

[56:00 - 56:01] John Wilbanks: Yeah, I talked- [crosstalk] I talked too much.

[56:01 - 56:01] Woman 2: [woman laughs]

[56:01 - 56:02] John Wilbanks: I apologize for that.

[56:02 - 56:05] Woman 2: It's wonderful. It's wonderful. What can we do?

[56:05 - 57:35] John Wilbanks: I think that's a great idea. So starting with, you know, giving... I mean, it's really simple. Give everything a name. Keep that name the same. Publish the name. Right. You know, I don't want to get biblical, but, you know, in the beginning, [audience laughs] you know, there was the name. And if it doesn't have a name, you can't find it. And if you can't find it, you can't do stuff with it. And so I think that one of the... given that we're in the United States and that most of the data is in the public domain, that's awesome, right? Like, I have such a harder time in Europe because I have to say, start with getting rid of the rights, right? Or Australia. Right. But we don't have that problem here. So I think starting by naming is an awesome way to go. Right? Just, you know, open up a set of servers where people can just load their data in, give it a DOI, right, encourage them somehow, some way to start attaching tags and keywords to that data. All right. That's a great tangible first step because then when this other stuff gets done, you can go in and rewire and connect everything without having to go get that, find that data and get it online. You know, the grad student that generated it will not have moved on, right? The data will not be on a thumb drive that has been lost. It'll be on the web. It'll be in your repository. It'll have a name. Right. That's actually one of the best ideas I've heard for just something to get started with. And if you do this right over the long term, it ought to be a way to compete for faculty. Right. Why should we come here? Because I get better data services here than anywhere else.

[57:41 - 58:02] Man 3: Okay, John, fantastic presentation. You alluded to how the distinct challenges with science, the grand science questions, as opposed to the Catch the Bus app. And then you also alluded to how a lot of the current apps we're seeing are like funded through advertisement. What's your recommendation for funding for the grand science challenges?

[58:02 - 01:00:24] John Wilbanks: I mean, if I had a good one, I wear a better suit. [audience laughs] You know, I think this is the sort of thing where, and I don't mean to be glib, but, I mean, there's not a lot yet. That money gets sort of bundled in into overhead as an assumption. You know, there's... I'd like to see a line item in grants that's like 3% for data sharing. But I think that's the sort of thing that would radically change the game. But, you know, we've been a part of multiple applications to the NSF for data sharing, archive and kinds of things. And there are all these, you know, five years and then when it's over, you'll have to be self-sustaining. You know, and that's not the approach we take to our highways, right? In 20 years, maybe they'll be self-sustaining, right? The internet, and you know, they didn't say in 75 this has got to be self-sustaining by 1981, right? But we've gotten into a culture where that's what we expect. And so I think part of what I try to do, part of the reason I fly all around and give these talks, and I spend a lot of time trying to convince people in Washington that they need to understand that and internalize this is an obligation of the grantor. Not to provide these as separate from research, but as integral as part of the research, because otherwise you're just creating these discrete knowledge units of papers. And while sharing those is a good thing. Sharing the raw materials that you paid for along the way is actually a pretty good investment policy as well. So the big thing I could say is that right now, if possible, write the data sharing into the grant as a line item. Try to give DOIs to those data sets and start participating in the emerging efforts to cite data, because then you can identify scientists who make good data, data that gets reused and reward them. And I think you can engage through URL and through spark in advocacy and in Washington to actually try to change the policy game, because I think... that's why I tried to divide it into business models, institutions and policy. I think that the funding aspect of this has to be driven by a change in thinking at the policy level. I'm not sure that it's something that any university, no matter how powerful, can change on their own. Thank you. [applause]

END TRANSCRIPTION