The Semantic Web for Publishers and Libraries

Michael Keller
Stanford University, michael.keller@stanford.edu

Michael Keller, "The Semantic Web for Publishers and Libraries" (2011). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/10.5703/1288284314870

The Semantic Web for Publishers and Libraries

Michael Keller, University Librarian, Director of Academic Information Resources, Founder/Publisher HighWire Press, Publisher Stanford University Press, Stanford University

Thank you. Good morning, everyone. So, before I start this talk I'd like to offer a few explanations and some thanks. First the thanks. This talk and the work behind it owe a tremendous amount to my colleague, Jerry Persons, who is Stanford's Chief Information Architect Emeritus. He continues to work on this particular domain with us and for us, and it's good because of what he's done.

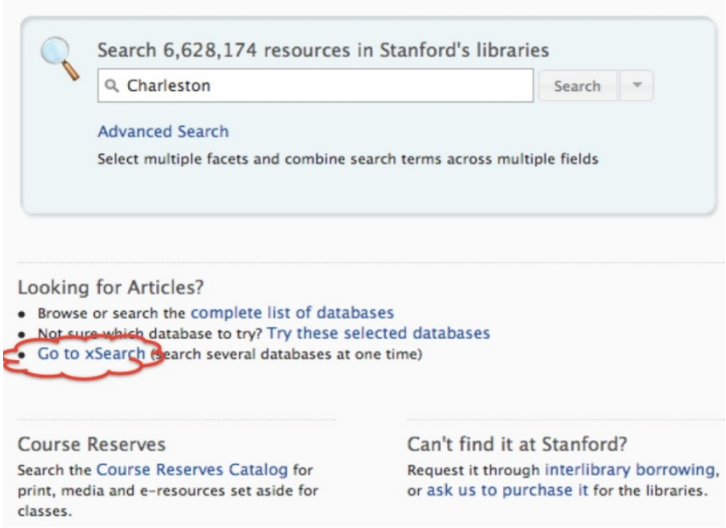
The explanation is that this is an introductory talk. Maybe I should ask right now how many of you are quite familiar with the principles of linked data and the semantic web? Please raise your hand. Perfect. So, the rest of you may learn something, and you may come up with lots of questions. There will be a lunch and learn session this afternoon with me, Jerry Persons and Rachel Frick from CLIR, and the DLF Program Officer, right in this room. So if you are burdened and there isn't enough time because Anthony is being very strict, or I talk too long, please come to that.

So, semantic web for libraries and publishers. I want to start with the problem set. What is it that we are working with here that is our concern? Frankly, the fact is we have way too many silos. We have red silos. We have concrete silos. We have blue silos. We have grain elevators that look like silos, only are bigger—you can imagine who those might be. We in the library and publishing trades have forced readers, some of whom are also authors, to search iteratively for information that they want, need, or think might exist in many different silos that use many different search engines, vocabularies, and forms of user interfaces. We do not make it easier for readers to discover what is locally available, what is more or less easy to access remotely, and everything that might be available.

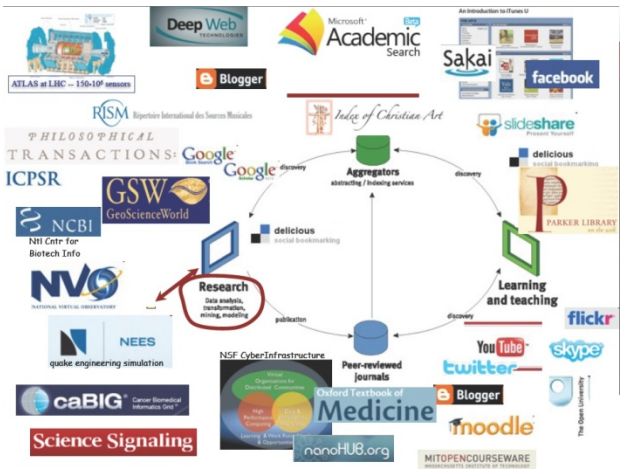
We give them better interfaces, including ones that permit refinement of a result, but these interfaces show our holdings more or less at the title level. An example of such an interface is the Search Works interface of Stanford based on Blacklight. Almost simultaneously we show the reader many other tools, some excellent in some ways, all of them good—because we select them, of course—and we suggest that our clients widen their search, to examine the literature more broadly.

However, no single tool is comprehensive. We routinely do not refer our clients to the web, at least not on our own websites. Our online public access catalogs (OPACS) don't refer them to the web either, except indirectly, when we have to go out on the web to look at an e-book or some e-information object or database that we've subscribed. While indices and abstracts refer our readers to articles and journals which we may have licensed, we rely on other services, such as SFX and the like, to provide the links to the titles which have been revealed through the search and the secondary publications. So neither are our OPACS, nor our secondary databases, directly referring to more than a tiny percentage of the vast collection of pages that is the World Wide Web. The web of course refers in fragmentary fashion to information resources we might—I emphasize might—have on hand for our readers. And the results of using those secondary publications or secondary databases, which are often very good, involve discovery tools and returns that involve relevance ranking determined in various ways. This provides us with different formats. This is a format from XSearch which is a locally branded product of Deep Web Technology, and various options for refinement that may or may not be different than our OPAC refinements if we have any. We therefore confuse our readership even more. Some of us provide our readers with lots of secondary databases, too many really, for all but a few who are forensic scholars.

DOI: http://dx.doi.org/10.5703/1288284314870



So, here is the Stanford interface. First you see XSearch down below, then we send the reader to select the database, we organize that by topic, and then we send the reader to the whole list. Selecting a database to search is something of an art. That's why we have good reference librarians and subject specialists. And notice once again that we do not offer the web as a search engine, as an option, and for good reasons. Nevertheless, the discoverable relevant information resources on the web apparently are not part of our repertoire in so far as these interfaces document. And, in the case of Stanford, we offer our readers the choice of 1,113 databases. This could take all day to sort through if really assiduous, I suppose. We somehow conspired—well, actually we haven't conspired; we're less than a conspiracy—but in some ways we have made the search for information objects very difficult. By "we," I mean librarians and publishers. We've just not had the tools, the methods, the vision, and yes, the gumption, to try something new.



The next slide shows a little teeny weenie miniscule portion of what's out there on the web that's relevant to the economic process of teaching and learning research that our folks have to sort through sometimes, mostly on their own. This picture multiplied by maybe 1 billion changes every day and gets more complicated every day, partly by the addition of new pages, partly by the addition of new subpages, and partly, frankly, by some sites just disappearing altogether. And the larger the number of websites indexed by Google or Bing or whatever search engine du jour, the more likely it is that the

relevance of the return will be less pointed or precisely matched to what the researcher thought he or she might find. So I return to my earlier statement: We've got way too many silos, in way too many places, with too many difficulties of determining what is in the silos and really with no way to get good returns on what's in some of the silos that might be relevant. This of course is the service that most of our students start with, particularly the younger and more naïve of them. It too, however, consists of silos.

Do you think one-size-fits-all in the Google world? It doesn't. Here are four of the principal silos: one for news, one for Google Books, one for Google Scholar, and one for Google Maps. That's on top of the Google main database. Google's main database is huge, growing, and changing all the time. These silos are very large, growing, and changing all the time, but, you can't look at each of these very easily except for through some clever interfaces they provided to these other silos. So given all the silos and search engines, our users—some of whom are authors, some of whom are teachers, many of whom are students some of whom are people on the street—need us to find a better way. We are wasting their time and we're not presenting them with information and information objects they need to have and they think might exist. Facts about the information objects we have acquired or leased, facts about books, articles, films, and so forth that we have published or licensed need to be found in the wild on the web. Ideally, we librarians and publishers will get the facts about what we have and what we're making public for fun or profit discoverable on the web.

So let's look at the problems a little bit. First of all there are too many stovepipe systems. Second of all, there is too little precision with inadequate recall. And third, we are too far removed from the World Wide Web.

Too many stovepipe systems: The landscape of discovery and access services is a shambles. I've shown you a slide to demonstrate that. It cannot be mapped in any logical way, not by us, who are supposed to be information professionals, and certainly not by the faculty and the students who must navigate this chaos. This state of affairs should not be a surprise. It grew up, as did Topsy. It just happened over the last 20, well, actually, over the last 150 years. There is too little precision with inadequate recall. Some of the problems are those various stovepipe systems. The dumbing down effects of federation often hinders explicit searches. And each interface has its own search refinements trick or tricks. There are numerous overlapping discovery paths hampering full recall. Most of the problem results from limitations in the design and execution of the infrastructure that supports discovery and access. In any given silo, that infrastructure may work very well for what is in the silo, but, it doesn't work very well across all the silos, and certainly not across the web.

A limiting factor is the problem of ambiguity. Most of our metadata uses a string of bytes to label a semantic entity. Semantic entity: people, places, things, events, ideas, objects. Discovery therefore is based on matching text labels, that is, on keyword searches. Discovery is not based on the meaning of the semantic entities, not based on the inherent meaning of whatever it is that has been labeled.

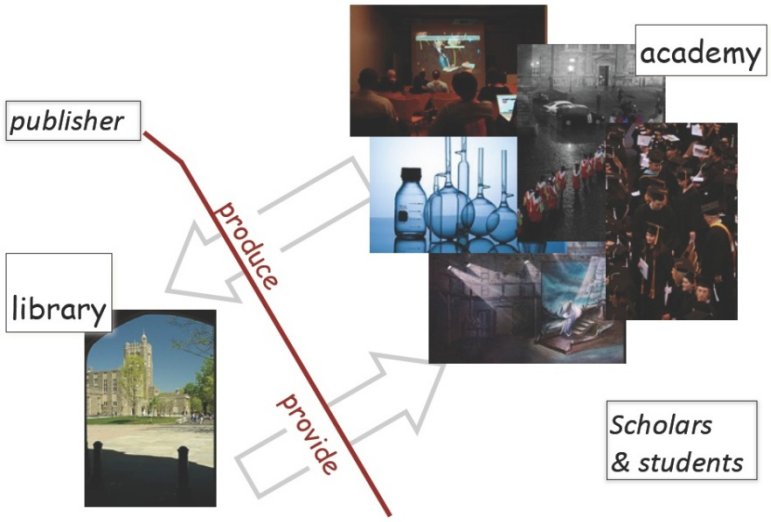
For libraries, our fix has been authority files. We have been really assiduous about developing these and they are excellent and we will make very good use of them in the linked data world. So authority files are authoritative strings, forms of strings, names, organizations, titles, places, events, topics and so forth, but, what about the case where no one-to-one relationship exists between a string of text label and the underlying semantic entity? What about the case one word has multiple meanings? Take for example the text string "Jaguar." All right, so here we have an example. We have the motorcar, the Jaguar, which introduces the SK series in 1996, the E-Jag between '61 and '74, and other ones coming out more recently even though it was once owned by Ford and for all I know may still be. There is hardware and software named Jaguar, there was an Atari videogame console called Jaguar, and the Macintosh OS 10.2 that was named Jaguar. In the music world, there was a heavy metal band formed in Bristol, England in 1979. A Fender electric guitar was named Jaguar, and there was a Jaguar Wright who was a singer based in Philadelphia. She was also a songwriter. In the military world, there is a type 140 Jaguar Class Fast Attack Craft Torpedo manufactured in Germany during World War II and

flown there. More recently there was an Anglo-French ground attack aircraft called Jaguar and there was, in the 1950s, a prototype XF 10 F Prototype swing-wing fighter made by Grummen on Long Island. Among the heroes, for those of you who are comic book fans or believe in fantasy in a big way, the Jaguar is a superhero in the Archie comics, and the DC Comics impact series also features a character called Jaguar. Of course in football there is the team in Jacksonville called the Jaguars. And now, finally, here is what I think of when I think of Jaguar: it is either a cat or car for me, but, here you have it. That is one illustration I think of the vocabulary of names proper and otherwise that create ambiguity.

The second limiting factor is the fact that we are evolving. We have evolved our systems to record a copy, a copy in our hands, particularly in the library world. So, most of the library metadata focuses on publication artifacts. We identify the responsibility for the creation of the artifact and we list topical headings. We describe it. For simple cases with an author that has very few titles, metadata translation, things work out pretty well. However, for authors with many titles, with many additions, things are much more difficult. So as complexity increases, precision and recall suffer dramatically, and we live in a very complicated world, as you know.

Here's a search that we did on the Socrates interface, the old interface of Stanford OPAC, on the terms Shakespeare and Hamlet—a very simple search. We get back 811 entities. Unflagging patience marks the task of flipping back and forth between hundreds of brief and full records to sort through the variances of the single entity. We have critical editions. We have 18th and 19th century collections of the plays**.** We have social and historical and literary answers. We have video and audio recordings of performances. We have reviews and indices of the same. We have treatments of stagecraft and costumes and music. We have the lives and work of others associated with the plays, that is, performances and directors. We have other art forms inspired by the plays. I've neglected to add here that we also have a collection of documents, information objects, people, and arguments that refute the idea that Shakespeare wrote anything, including Hamlet.

We're too far removed from the World Wide Web. Together our metadata collections make up a big chunk of the dark web, the web that is not indexed. It is clear that visibility on the web promotes dramatic increases in discovery and access. So if you take a look at the traffic against the Flickr images from the Library of Congress and the Smithsonian you'll see a lot of traffic. When in 2002 Google began to acquire an index of articles published through HighWire services, we had a dramatic increase in the amount of traffic back then it has since increased. So this state of affairs is very well known.



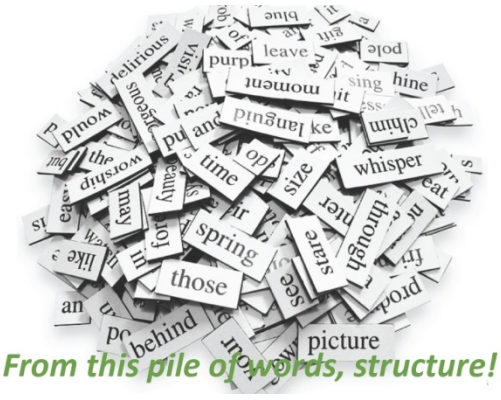
What is our working environment, what are we dealing with here? Take a look at this schematic to see the ecosystem in which publishers, libraries, students and scholars are involved. Now this is very

simple, I give you, very simple. But you get the point. We have consumers and producers in the upper right-hand corner. We have the publishers and intermediary taking some of those products and turning them into published works which they sell or which we somehow acquire in our libraries which we then feed back to the students and scholars. And, another piece of our ecosystem has to do with the network that we communicate on. Some years ago, many years ago, there was the Internet. There wasn't much e-discovery or analytical communication going through that. But, we had a whole bunch of prophets; three of the most important were Vannevar Bush, in the mid-50s, and Ted Nelson and Doug Engelbart, who predicted what the Internet could become. And then thanks to another prophet, Tim Berners-Lee, the Internet became a web of pages of information.

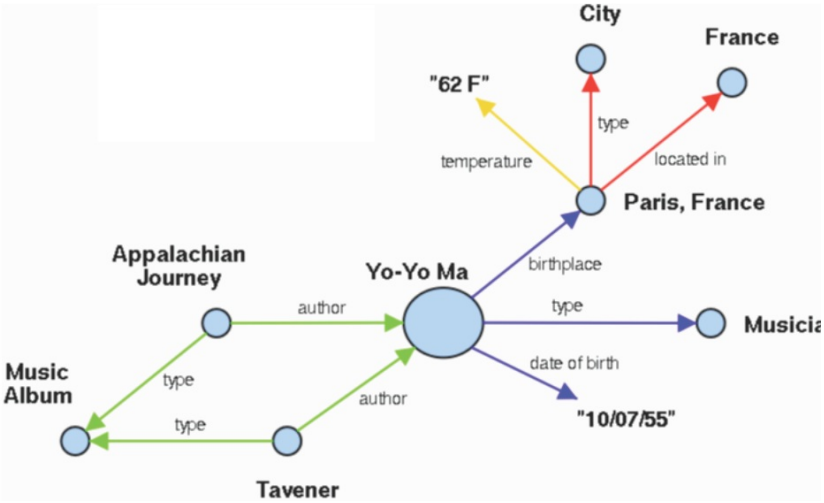
Scholarly journal publishers and some librarians realized early on that there were functional advantages to scholarship and to publishing in the web of pages. Yahoo, Google, and others realized that mining the web of pages by words off the pages could make a rapidly growing web of pages reveal more through indexing and cataloging. As a matter of fact, indexing won out, as we now know, over cataloging. The web of data is the next big thing in discovering relevant information objects and the next big thing in empowering individuals, communities and industries and making better use of information that they or others create. What distinguishes this web of pages, this linked data environment from the web of pages is the principal of identifying entities, virtual and real, like statements of relationships which are therefore descriptions of machinery before.

We are calling this next phase the linked data phase because it is entirely dependent upon statements of relationship and descriptions. But this phase is only a precursor to something even more complex and certainly more difficult to engineer. And that phase is the semantic web, which in theory will allow the machinery to build relationships and descriptions, to interoperate with themselves to satisfy requirements, requirements made by another system, requirements made by person, albeit without constant interaction with the demanding body, whether it's a machine or a person. In short, in the semantic web the machines will understand meaning and presumably act upon it. That's a scary thought.

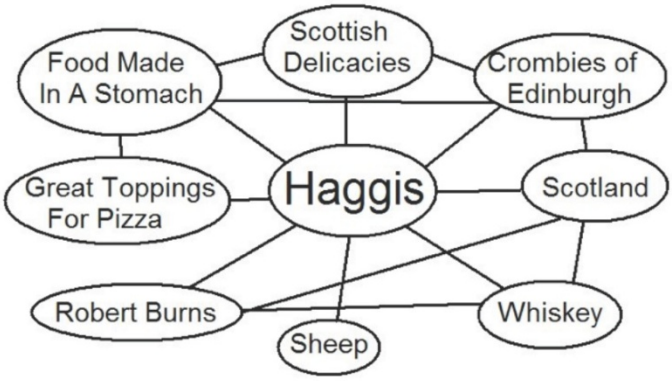
So what are the tools that are going to get us there? How do we work to alleviate our problems as information professionals, as librarians and publishers. Here's the recipe: we identify people, places, things, events, and other entities including ideas, embedded in the knowledge resources that a research university consumes and produces. We tie those facts together with names and connections. We publish those relationships as crawlable links on the web. Crawlable and open for anyone to use. And we build and use applications that support discovery via the web of data. Some of those apps I can describe for you in primitive form today; I'll show them to you.



Here's a pile of words representing, in a very small way, all of the words on the web that most search engines constantly use and constantly index. Good search engines can do a lot with this pile, but the search engines create a perception of relationships based on other factors such as the number of links containing words of interest or the traffic to a site. And from this pile of words, actually from the pile of webpages containing the words, we are going to build this linked environment. The structure of the new environment will be based on the meaning of the relationships.



Here is an example, a very simple example, of how these relationships can describe a person. This is a graph of Yo-Yo Ma the great cellist you'll see his big blue circle in the center there. This is only a small tiny bit of relationships that he has. So he was born in 1955; he is a musician; he loves the city of Paris, which is in a certain country where the temperature is a certain temperature. He's made a recording entitled *Appalachian Journal,* which is a music album. It features, among others things, the music of John Tavener. This is a graph that demonstrates how relations begin to define the elements on this page. Each of these elements has a relationship through one means or another, through one hop or another to all the others.

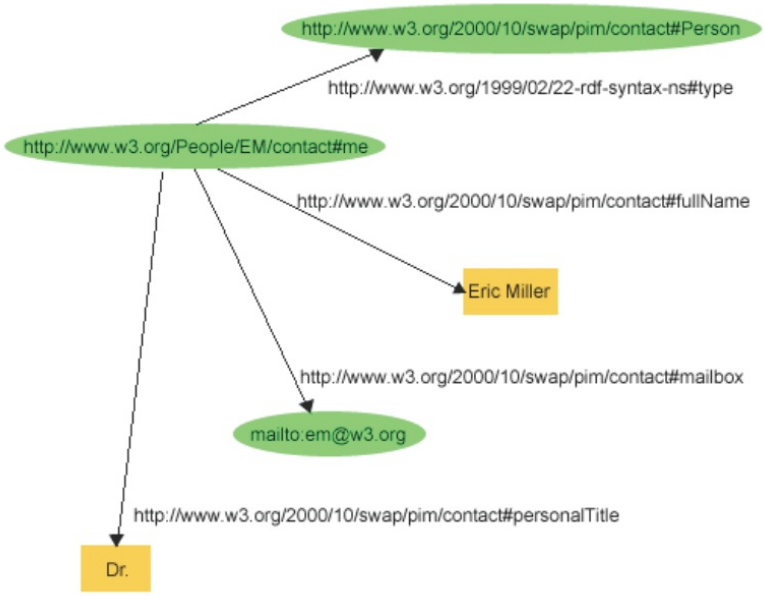


Linked Data Web

Here is another one, this is a silly one, and I have to confess there is one aspect of this I really don't understand, but someone from Scotland will have to elucidate. So, this place is haggis in the middle of the picture. Absolutely haggis is a food made in a stomach, literally. It's a Scottish delicacy, so they say. Crombies of Edinburgh manufacture or make it, and it's Scottish. It involves a certain amount of whiskey, I presume, before, during, and after. It involves sheep. Robert Burns has apparently written

about haggis. The “Great Toppings for Pizza” I don't get. I think there is some oatmeal involved so I'm having trouble putting the oatmeal on the pizza.

Okay now some geek talk: RDF triples and URI's. Resource Description Framework always expresses a simple sentence: subject, object, predicate, is a way to describe objects or even ideas on the web. An object or an idea may have many RDF triples describing it because everyone of us have many different relationships and there are many different ways to describe us depending on where we are, who we are, what we are doing, and so forth. And, as I said, objects or ideas need not exist on the web. URI’s: Uniform Resource Identifiers. Like URLs, only stable and steady. These allow machine interaction among Web Objects provided with various and tactical schemes and protocols used to construct to URI’s. So there is a vocabulary, there is a way of expressing URI's that is well-known and being built-up principally on the World Wide Web Consortium in Switzerland with our support. We need at least three of these to support an RDF: subject, predicate, object.



Here is a graph of URI's with an RDF. The RDF is Dr. Eric Miller and the green bits are the pointers and the unhighlighted bits are the syntactical ways of expressing these elements, the elements of this sentence: Who is Dr. Eric Miller? Where is Dr. Eric Miller? Here are the linked data principles: use RDF’s as names of things; use URI's so that people can look up those names. And when someone looks up a URI, it provides useful, actionable RDF information from URI’s and include RDF statements that lead to other URI’s so that the reader can discover related things.

Back to the problem of library metadata. Our metadata standards are closed. We have spent innumerable hours over 70 years devising these standards, modifying them, and so forth. It is a big industry. But they're closed. Passive metadata is searchable by word, by string, but it is in the silos. It's readable, it's not actionable, it’s passive. The search results are refinable, but they are final. They don't take you another step; you can’t go beyond the search results of your OPAC or very many of the segment publishers situations.

Here is a comparison. We’re going to spend a little time on the right column there. Semantic metadata is open, or should be, it is dynamic, it conceptualized, it is living, it’s actionable, it's not passive, it exists in an environment—an ecology—of lots of these things. It is in the wild, ideally, it can be used. It’s interactive and responsive; it can take you places; you can do things with it. You can resolve it with words; we can look at it with graphs, or both. It can lead to other queries and other views. So my plea to all of you and to the world wide web of libraries and

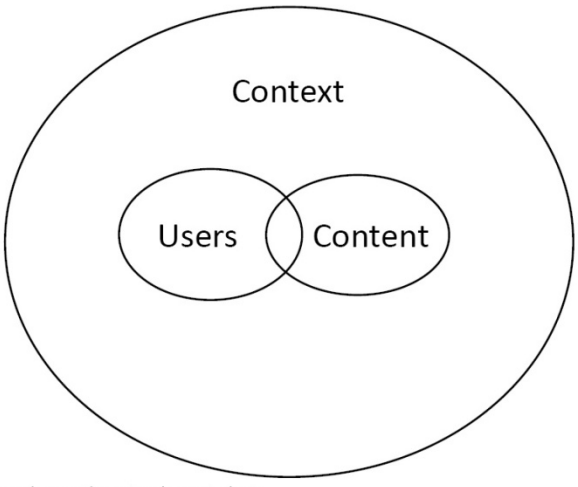
publishers is to make library bibliographic facts into RDF's and URI's, release them into the wild, and make library linked data open—usable by everyone.

What about publishers? Why would publishers be interested in this? Well, publishers should be interested in aggregation. Aggregating their content in their own realms and allowing aggregation of their content within other realms. They could aggregate information beyond their publications, beyond articles and books, to information about conferences that are relevant to the subject of an article or a book, to career building and employment opportunities, to collaborative communities, to commercial and other services, to advertisers—who support research, ostensibly, with specific source materials, processing, and trials—and to produce productive relationships with others. Publishers should want to provide actionable and constantly updated links in support of scholars, teachers, learners, and those in the academic publishing trade. And they should be interested in providing compelling tie-in users to the publishers themselves.

Here are some of the entities that are already committed to making accomplishments in the sphere. There are a lot of them and this is only a small selection: the Associated Press, the United States Department of Defense, C|Net, the Library of Congress, the British Library, Google, Wolfram, Thomson Reuters, Hearst Interactive Media, Novartis, PLoS One, The Guardian, Elsevier, Pearson, the British Museum London, BBC, HighWire, Merck, and Astra Zeneca.

I want to specifically mention a few. The British Library not many months ago leased the entire British National Bibliography in RDF and URI's. The entire British National Library. This is a tremendous contribution. The Library of Congress has released the Library of Congress subject headings and the name authority files as RDF's and URI's in the wild. And the subject headings have links to (28:44 of the video) Aggrevot, Rummelo DMB, the GLN Subject Thesaurus and the National Agricultural Library subject index. Every personal and corporate entry in the LC name authority file linked to the virtual international authority file, basically OCLC. The VIF is not yet open; it has not yet produced RDF's and URI's that I know of in the wild.

Very significantly, about 18 months ago, the New York Times released into the wild all 500,000, and growing, of its index terms for use by anyone. That is tremendous. That is a whole other vocabulary outside of the ones we usually use. For publishers and libraries content is king. Although none of us should neglect services: services to our readers, our authors, and our institutions. However, if users cannot find content in their own context, there is a problem. Therefore, if you understand users to be readers, authors, teachers, and students, the following Venn diagram suggests the overlaps.



Users = readers, authors, teachers, students

Now, I believe publishers must make their content visible. Indeed, it's an imperative, because if the published content is invisible there is no benefit in tangible or intangible form to the author and certainly no benefit to the publisher. This is a PLoS article that was published in 2009 in their journal on Neglected Tropical Diseases. It was symantesized by David Shosen and a few others at Oxford, and all those highlighted elements have information behind them. This, however, is not actionable; this was all hand built. It took 10 men weeks to build it. It is, I believe, possible that we will be seeing more of these as we do a lot of tagging, as the publishers come up with better ways for semantics to be installed using RDF's and URI's. So, eventually you'll be able to see lots of these with links from the terms into information resources explaining to them. I've already mentioned aggregation, and I couldn't resist putting this slide in front of you. But, for libraries and publishers, aggregation is very important, and I emphasize, as this slide emphasizes, the multiple different forms that information objects might turn out to be in a really good aggregation. It doesn't all have to be articles that could be documentaries. It could be sounds, it could be webpages, it could be printed and published things.

So, are we still confused and lost? Do we still have this problem of ambiguity? Well, yeah we do sort of, but there is a way out of it, and this sign in the upper left-hand corner—although it is not readable to most of you—is actually disambiguating a direction. And the point I'm making with this slide is that in the RDF, URI, or in the linked data world, there are very easy ways to make very arcane languages readable. The arcane language in this part of the slide is Irish.

So what is the web and data progress? In 2007 these circles represented the agencies that were broadcasting, publishing URI's and RDF's. This is that same environment in 2011. Up here we have hundreds of millions of URI's and RDF's occupying gigabytes of content. Now we have hundreds of billions, going to trillions, of these entities out there. Fortunately they don't take up that much space because they are very short. So, there's some encouragement.

Here is the linked open data value proposition that was developed at a workshop we did at Stanford in late June. Linked open data puts information where people are looking for it on the web. Linked open data can expand discoverability of our content. Linked open data opens opportunities for creative innovation, endeavoural scholarship, and participation. It allows for open and continuous improvement of data and creates a store of machine actionable data on which improved surfaces can be built. Library link to open data might facilitate the breakdown of the tyranny of domain silos. Linked open data can also provide direct access to data in ways that are not currently possible, as well as provide unanticipated benefits that will emerge later as the source expands exponentially. Here is a slide which shows a linked open data application in action.

It's from Freebase, a Google company now, and it's based on bibliographic facts from Stanford and web resources. It is about Stephen Jay Gould. You saw the editions of *The Panda's Thumb.* Now you see the description of the book. Now you see excerpts from the book. A lot of them. Now you see a couple of reviews of the book. All of this is being created on the fly; it is not hardwired using RDF's and URI's. Here are the RDF's, and you see there are a whole bunch of them there, that have been built, developed algorithmically for the site, sampling them from here and there. Now we go to look at Stephen Jay Gould. We're looking at the *Panda's Thumb* site. Now we're going to take a look at the site that is associated with the RDF Stephen Jay Gould. You'll see a wiki biography of Steve; you'll see a list of books, some of which are readable on the web, a lot of which are underlined. You'll see the same environment, its papers, and some of them are highlighted because they're machine-readable. You see a video, this is where the sound comes up, I hope (video begins playing in background).

We'll look at some quotes from Steve that are from books and articles, reviews that he's written—all of this assembled on-the-fly using this linked data environment that was built at Freebase. I think we're going to look back at the papers because I need to show you something about how the papers function can work. These are people who cited this particular article and you can go to the next tab over and look at the citations. Now we are going to look for Dawkins, Richard Dawkins. It takes a little while for the machine to think, this is not logged, by the way, this

is a movie. Here we start on the Dawkins slides. All of this is done with linked data, all of it done with bibliographic facts from Stanford and web resources of various kinds. The BNF, the Bibliotheque Nationale de France, has created another interesting example using only data that they control, only bibliographic information they control and digitized content from Gallica and another movie. So now we're going to look at Victor Hugo, the complicated author, for a variety of reasons, a very prolific author. You can see his pseudonyms; you can see the sources of the information about Victor Hugo and his output; you can see his works, lots of them, a whole lot of them. On the right where it says "Visualiser" it means this is where you can go to read the title in question or the addition in question. We went to Les Mis and we're going to look at the books, enormous number of editions of Les Mis, hundreds actually, but also their translations. They are as you know the brechti for operas and for musical productions. Les Miserables appears in anthologies, all of that indexed in this site.

On Monday, Halloween, Library of Congress announces a bibliographic framework for the digital age. A new bibliographic framework project will be focused on the web environment; linked data principles and mechanisms and the resource description framework as a basic data model. They have put down the notion that we're moving from MARC to linked data; it is going to happen. The value proposition, which is also from that Stanford conference, would promote the following practices. This is 25 people gathered at Stanford from a variety of institutions: "We want to publish data on the web for discovery and use rather than preserving it in the dark more or less unreachable archives that are often proprietary and profit driven. We want to continuously improve data and linked data rather than wait to publish perfect data. We want to structure data semantically rather than preparing flat unstructured data. We want to collaborate rather than working alone we adopt web standards rather than domain specific ones. We use open commonly understood licenses rather than closed or local licenses."

This is where we started when we went to the World Wide Web. This is the social web which floats on the World Wide Web but we must pay attention to it in our field. I remind you of what the linked data web looks like, what it is in terms of relationships, and how relationships describe meaning. We're headed to this; we're headed to the semantic web. A couple of big ideas that accompany these notions: The first is the ubiquitous computing that is essential and makes it possible for lots of players, people, and institutions around the world to participate. The mobile communications part of that ubiquity is very important, as it allows people to use the linked data web wherever they happen to be. So that is the way that the world is progressing. This is what we don't want any more of.