E-Everything: Putting It All Together

Peter McCracken
ShipIndex.org, peter@shipindex.com

Follow this and additional works at: <https://docs.lib.purdue.edu/charleston>

An indexed, print copy of the Proceedings is also available for purchase at:

<http://www.thepress.purdue.edu/series/charleston>.

You may also be interested in the new series, Charleston Insights in Library, Archival, and Information Sciences. Find out more at: <http://www.thepress.purdue.edu/series/charleston-insights-library-archival-and-information-sciences>.

Peter McCracken, "E-Everything: Putting It All Together" (2010). *Proceedings of the Charleston Library Conference*.

<http://dx.doi.org/10.5703/1288284314817>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

E-EVERYTHING: PUTTING IT ALL TOGETHER KEYNOTE

Peter McCracken, (peter@shipindex.com) - Co-Founder & Director of Content and Business Development, ShipIndex.org

The “E-Everything” preconference included the following speakers: **Audrey Powers**, University of South Florida; **Sue Polanka**, Wright State University Libraries; **Peter McCracken**, Serial Solutions and ShipIndex; **Jason Price**, Claremont College Library; **Michael Gorrell**, EBSCO; **James Mouw**, University of Chicago; **Lisa Carlucci Thomas** – Southern Connecticut State University; **Stephen Rhind-Tutt**, Alexander Street Press; **Emilie Delquie**, Publishers Communication Group; **Cory Tucker**, University of Nevada, La Vegas; **Anh Bui**, HighWire Press

Preparing for today has been an interesting, stressful, and enjoyable project. As I looked at the roster of speakers who will be following me today, I must admit I felt a fair bit of trepidation; I can only hope that at least a majority of you all will find my comments interesting, and perhaps even thought-provoking. I understand that there’s a nearly equal balance between librarians and publishers in the audience today, and I think that’s a great thing. Certainly, that’s one of the best benefits of the Charleston conference, and I am so thrilled to be a co-conspirator in spurring today’s conversation and perhaps further thoughts and ideas over the course of the next four days.

As we start this preconference, I think it would be very useful to consider how far we’ve come in presenting electronic resources to our patrons, and offer some thoughts about where we’re headed. I’ve developed a proposal that I feel would be a very useful and improved way of delivering content derived from electronic resources to patrons. I want to present it to you today, and hopefully encourage some discussion about what might work and what might not. I have absolutely no commercial interest in this idea; as far as I’m concerned, anyone in the room can take it and run with it, as long as they include ShipIndex.org in its final result!

My thoughts today are, at best, moderately-educated guesses, and you very well may disagree with some or all of them. Most likely, you’ll have some better ideas about what patrons want to see or better ideas about how to implement the vision I’m trying to present. That said, I think that my ideas are generally informed by what I think will work, with only a smattering of crazy-futurification-ideation to shape it. There is a bit of that, though, and in fact, I’m not sure how practical today’s ideas are. But perhaps all they’re waiting for are some technological enhancements – or, more likely, the correction of a lack of knowledge on my part of already-existing technological tools to make them work.

I want to look at where we’ve come from, mostly so we can remember how far we’ve already traveled. Then consider where we are now, and take a few shots at where we’re

headed. I look forward to thoughts on how you all disagree or agree with me. The crowd is certainly smarter than the individual, but only when it speaks up.

So, where have we come from? I can't decide if it's been a long time since I started as a librarian, or not. On the one hand, we certainly had internet access and I had put web pages online before I started working in an academic library -- it's not like the card catalog was getting a lot of use when I started working as a librarian. On the other hand, it was 15 years ago. I did participate in the weeding of the physical catalog myself (we cleared out cards representing items that were in the online catalog, dramatically reducing the size of the remaining card catalog) and I still have a collection of catalog cards at home. I find them really useful for writing notes for my first-grader, to remind him when he's supposed to go to a friend's house after school. Also, I was looking at the website for an individual who graduated from the same library school as me, but 10 years later, and I felt like I was looking at someone from a completely different generation. And, I suppose, I was. I'm sure all the catalog cards were gone by the time he got there. Clearly, there have been a lot of changes in how people view and learn about librarianship and how they apply that different knowledge. But not everyone currently working in a library graduated from library school in the last six or seven years, so of course the variation is really the amount to which we are each adapting and learning as we go in our libraries each day.

What I recall about electronic resources from my first job as a reference librarian, when I was working at East Carolina University, from 1997 to 1999, was the growing domination of CD-ROMs. We had multiple machines on which one would use hundreds of different specialty CD-ROMs. I remember finding, buying, and installing the CD for Cornell's Hospitality database, which was relevant for some program at the University. A maritime history student was desperately awaiting the arrival of Cambridge University Press's Trans-Atlantic Slave Trade CD-ROM database for her thesis research. But most prominently I remember the big metal stacks of ProQuest CD-ROMs in funky plastic cases. It's fun to think back to the times when our reference departments had so many CDs in so many places, they looked like an AOL marketing distribution facility. And I don't mean this as a knock at all on the technology we were using -- in fact, being more of a pragmatist than a futurist, I was very concerned when ProQuest wanted us to convert our access from their CD-ROM technology to internet-based access, because our internet connection was often flaky, at best. I definitely worried about the quality of service we'd be providing to our patrons.

But, clearly, there's no news here to speak of - the pace of change in using electronic resources in libraries is as dramatic as in any other field, other than perhaps barbershops. As I said, we used CD-ROMs as the primary method of accessing electronic resources. It was a bit klunky, but it could have been much worse -- and of course we didn't have to worry about whether we could access the internet at any given

moment. Adding new content certainly wasn't instantaneous – we had to receive new CD-ROMs every month, and load new indexes from CD on to each and every workstation we had – but it did provide vastly superior access to resources over print-only indexes or other tools. When we did shift to internet-based access (and our internet access was reliable), of course it was another significant improvement. In a similar dramatic shift, all the data from that 1999 Trans-Atlantic Slave Trade database has been released online, along with enormous amounts of additional data, at the website SlaveVoyages.org. Online, it's all available for free, underwritten by support from Emory University. This is, certainly, a dramatic improvement in how people can access and use that valuable database.

When I went to the University of Washington, in 1999, they had just given up trying to have their own interface to all databases. It was a text-based interface, if I remember correctly. They had done what they could to manage a single interface for users, to increase students' ease at accessing those disparate resources and interfaces. This was an honorable goal, but the proliferation of online resources, even then, made it impossible, or at least impractical. As I think about it, and look to the future, though, I think there's an opportunity that we could actually get back to that.

And now, where are we? Obviously, we've seen an enormous growth in the availability of electronic resources. Most patrons prefer electronic access for obvious reasons, and I feel certain that Serials Solutions, and other electronic resource management companies like it, have dramatically simplified the ways in which patrons use electronic resources and librarians manage those resources. It is undeniably true that it is much easier to find information now than it was just 10 years ago, to say nothing of two or three or more decades ago. So, this is where we've come from, and this is where we are right now. It's clear that technology has improved access to information, and that electronic resources make information much more accessible than was previously the case. It's worth emphasizing that the online catalog, for all its limitations, is clearly superior to the card catalog. But at the same time, poor implementation of technology does a great disservice to library patrons and their needs to find information. The technology we offer today, for accessing electronic resources, is often woefully inadequate when compared to other technology, already in use, in other fields.

I feel that, overall, librarians are very good at thinking about the service they're providing to patrons, and constantly working to improve that service. We should all commend ourselves for that, but at the same time we must be sure that we don't stop trying to improve. So, in that vein, I want to offer some ideas for further improvement of the online resources experience.

In Kevin Maney's recent book, *Trade-Off: Why Some Things Catch On, and Others Don't*, which I should point out that I found by browsing the new book shelf at the public

library, he argues that most organizations succeed by focusing on either “fidelity” or “convenience”. “Fidelity” describes the quality of a service or product. The example he cites, and I like best, is of attending a U2 concert – it’s incredibly expensive and inconvenient, the time and location are on their terms, not yours; it takes up a ton of time; but the experience is amazing. “Convenience” describes the ease of obtaining or using a service or product, including the impact of pricing. An MP3 recording of a U2 song is very easy to obtain and share – it’s not as high quality as a CD recording, but it’s plenty convenient enough for nearly everyone.

There are a number of items that get lost in what he (unfortunately) calls the “fidelity belly”. A good example is Blu-Ray DVD players: they don’t have a sufficiently higher level of fidelity to justify the reduced level of convenience (in this case, “convenience” primarily relates to cost, though it is also harder to obtain Blu-Ray DVDs). Another example he uses is the Kindle: I think it is more convenient to acquire a book via a Kindle than via traditional means, but then using that new book is not nearly as convenient in many ways – have you ever tried to flip through a book on a Kindle to find a specific passage? Forget it. And fidelity – that is, the quality of the experience? The iPad has clearly popped that bubble.

In the library, we have multiple paths to both fidelity and convenience, and one can pursue them in different ways in the same institution. An interaction with a reference librarian should be about fidelity: the librarian providing resources and assistance that the person couldn’t otherwise find, despite the inconvenience of coming to the library and other related issues. Circulation should be about convenience: first, in making the resources easy to find through the catalog; second, in making it as easy as possible to get the resource home. The hours might not be quite as good as the Barnes & Noble that hasn’t yet closed, but the cost certainly is.

We must always be working toward the best fidelity or convenience; that endgoal is always moving, but we always need to be moving toward it, as well; and in libraries we have done that, as we’ve gone from print to CD-ROM to online access.

The current OPAC interface directly addresses the question of “convenience”. Trying to find a book in an online catalog is hardly convenient; in many cases, the most convenient way of doing it is to search for a book on Amazon, then see if the library owns it. That’s no good at all. I’ve got plenty of examples of the challenges associated with the OPAC – as do you, I imagine – and I’m convinced that far too many people are pushed away from using – and then supporting – public libraries because of those lousy OPACs.

Another challenge to both fidelity and convenience is the proliferation of electronic resources. In and of themselves, of course more is better. But the more resources there

are, the harder it is for a library to choose which of those resources to buy, and which free ones deserve inclusion in the library's database directory. How can any one librarian know which resources they should know about? There are over 16,000 entries in the latest (online) edition of Balay's Guide to Reference Sources; even after they've gone from a print to an online format, I expect the editors must spend quite a lot of time and energy simply deciding which resources to not include. Proliferation of resources has been difficult enough over the past decade; even when libraries take a more-enlightened approach to database organization than an alphabetical listing, organizing stuff by broad subject is challenging – and inter-disciplinary work is, obviously, the norm these days.

One solution to these problems would be the wholesale application of linked data and the semantic web to scholarly electronic resources. From this, we could create a truly single source for searching. Like any semantic web solution, this would require much greater interoperability among all users, and that's not an easy thing. Imagine, however, what the end result might look like, and consider if it remains a goal worth achieving. Consider doing a search and getting information back from a database you didn't know you'd searched. In fact, you didn't know the database existed before you started the search, and – more importantly – you didn't know the database existed after you completed the search. Discovery tools, such as Summon from Serials Solutions, EBSCO Discovery Service, Ex Libris' Primo Central, are a start. But even then, of course, the discovery tools tell you that your results are coming from a specific database, and – most importantly, for this proposal – you're only able to access the results from the databases that you subscribe to.

The discovery tool can return results you can't access – that is, the database will tell you that other resources, which you cannot access, are available. I guess there may be some good in that, but I'm not quite sure what. Some patrons could submit requests via ILL – primarily graduate students, faculty, and some advanced searchers – but in most cases, if it's not available right now, it's probably not that useful. The “convenience” factor outweighs the “fidelity” factor for many library users.

But if you could get past this, it could be a liberating experience, particularly for small schools. I live outside Ithaca, New York. There are three institutions of higher education in Ithaca: Cornell University, Ithaca College, and Tompkins-Cortland Community College, universally known as TC3. If each institution subscribed to the same discovery tool, an identical search done today at each institution would return very different results, even though they're accessing the same underlying index. This is, obviously, because they subscribe to different databases; Cornell subscribes to many more databases than Ithaca College or TC3. What would be the impact if this changed – radically – and the same search at each institution returned the same results?

In my proposed paradigm, you enter a search term or a set of terms. You're shown a faceted search result that helps you narrow down what you're interested in: when you type in "Magellan", you use the faceted results to choose from the mutual fund, the explorer, the geographic location, or the spacecraft. Faceted results are critical, because we are dealing with such an enormous range of resources. And once the data has been presented in a faceted manner, and the user has narrowed down her search to the point where she wants to see results, those results are all pulled from multiple locations – and most importantly, not just the databases to which her library subscribes.

Similarly, later, while searching through results, the individual can find more about a phrase, or a topic of interest, directly from within the database. Imagine that you didn't know that a niche database existed, but when you needed its expertise, you got it, via the standard interface you're already familiar with. Again, results come not from just the databases to which the library subscribes, but from all databases that can be included in the results set. I'll address some thoughts about how this is defined as I move forward today.

This solution would focus on both fidelity and convenience – fidelity, in the sense that the best data available would be included in the results' contents. It's not just what's freely available online, but what's available in subscription databases – and, again, including the subscription databases to which a library does not subscribe. The quality of the results is unrivalled. It's the best data available, and because it's clearly defined via linked data, it's easily findable. Convenience is also obviously met; there's no way to compile this data as efficiently, because you'd have to go to each database to do the research, decide if the results are worth including in the record set, and then bring them all together. And, you'd need to know where to go to do the research – including the databases that you don't know exist. Convenience also reflects the actual cost associated with acquiring the data, and I'll address that soon, as well.

Making such a process and product work would require input from a lot of different groups. I think it would be useful to define (at least in the way that I see them) who and what is involved here. First is what I am calling the primary vendor. This is the organization that provides the initial interface; they sell that interface, along with whatever full-text data they already offer. In addition, they manage payment issues, going off to collect and then display the data from different databases. They compete with others in this arena on their interfaces, their back-end services to librarians, their interactions with secondary vendors, and cost.

The database particularly shines when one is using it from within another article or entry, returned from the database. That experience in the interface might look something like the following: A patron has done a search and has found an article that mentions Frank Sinatra, the USS Gerald Ford, various fish, philosophy, and classical

literature. This is a seriously multi-disciplinary article. The user wants to learn more about one of these items. She'd click a "turn on links" button and then select the term she wants to learn more about. She clicks on "Oncorhynchus tshawytscha". Since the article contains only this single reference to a scientific species name – but uses a scientific name rather than just the common term "Chinook salmon", the decision engine selects resources from several general-subject animal resources, such as Grzimek's Animal Life Online, Aquatic Sciences and Fisheries Abstracts, the Regulatory Fish Encyclopedia, or perhaps brings back something from Public Affairs Information Service. But it doesn't just provide links to those sites; instead, it presents data from those sites in its own format and layout – the one the patron is already familiar with. The ancillary features of the primary vendor, such as the ability to email data, or compile bibliographies, remain available. The data from those other sources appears in the primary source, and it appears as part of it.

Another service provided by the primary vendor is overall analysis of the databases in use, similar to what discovery layer vendors are doing today. However, instead of just describing usage on the databases to which a given library subscribes, they'd also provide quite valuable information from the databases to which the library does not subscribe. By looking at how often results from unsubscribed databases appear in results sets, the library could make smarter decisions about selecting databases that their patrons are using, but the library doesn't subscribe to. Of course, these results don't include usage for databases when patrons go directly to them, but they would provide comparable usage across subscribed and unsubscribed databases from within a general-purpose interface.

For patrons at the largest institutions, where they already subscribe to these resources, there's no additional cost to them – they've obviously already paid their subscriptions to Grzimek and Aquatic Sciences & Fisheries Abstracts online. Perhaps they have access to most of the databases I've previously mentioned, but there are one or two that they don't purchase for some reason. They're now displaying content from essentially every database that might be relevant to their patrons.

For those at smaller institutions, the data remains available, but at a cost. Instead of buying buffet access to the database whenever they want, the library has a la carte access to results within the primary vendor's interface. The primary vendor manages the backend financial relationship, through their pricing agreements with the library and with the secondary vendor. Since the library doesn't have a subscription to the database, their patrons cannot go directly to it, and use just it. But they still have access to it, and through an interface they already know.

The financial aspects are important; this is an area that I think cannot be overlooked. It's great to sketch out a system that would be ideal for patrons, but if it can't keep itself

alive financially, then it's really not worth pursuing – until one can develop a financial and practical system that will support it. So, I think it is worth exploring the system, and seeing if it will work, both in a way that serves individual users, and also can keep itself financially afloat.

Secondary vendors are those who provide useable data in a semantic format. Their content is delivered to the patron via the primary vendor's interface. There is certainly a place for the secondary vendor to brand its product, but it is limited in this instance. The patron should know where the data came from, but this doesn't need to be announced from the treetops. And, the secondary vendor still offers access via its own proprietary interface to those who subscribe directly to its service.

The library manages how the resource is presented, on behalf of its patrons. It also provides significant branding, as the library should always receive credit for the content being offered to the patron. The library can control, via the primary vendor's back-end interface, how many and which databases offer results to their patrons.

A key to making this work is ensuring truly effective communication between the primary and secondary vendors (or, rather, between their computers), and that requires interoperability, and possibly standards. I'm not sure if we need standards to make this work, or if we can agree on enough issues that we can create taxonomies – textual structures that define all the possible data points that might be used among and between subject areas – that essentially agree with each other, at least to the extent that they will operate effectively. The definition of these “upper ontologies”, as they are called, can be a very political and challenging issue. There are plenty of standards that can probably be applied – as the saying goes, the nice thing about standards are that there are so many to choose from.

Numerous subject-specific structures already exist, and I'll touch on those in a moment. The challenge, it seems to me, is finding a way to bring all those subjects together, under an upper ontology that can be accepted by all parties, and over all of which someone can exercise control.

At the specific subject level, concepts of linked data and the semantic web work well. By creating a semantic, or unambiguous, or specific, data structure and then accurately tagging and defining resources so they make the greatest use of that data structure, nuggets of information and content can be easily located. By using linked data – defining in a more permanent way where on the web information about a resource exists – such data can be quickly accessed. Together, they can make it much easier to locate and then collate information in the cloud, whether it is in free or subscription databases.

A lot of different structures for managing linked data do exist; one of the largest being that at LinkedData.org. LinkedData.org incorporates data from DBpedia, which contains nearly all the others on the site, and is based on data drawn from Wikipedia; US Census data; and several others. The Open Library has records for several million books, and other databases have data about music, geography, chemistry, etc. There's a "Poképédia" (and a "Pokedex"), the Fishes of Texas, and one just about the Goodwin family. Arrows shown between data sets suggest connections, but I'll admit that I'm unclear how effective those links are. What we want to see are links between every data set that contains overlapping content or concepts. But there's so much more that's not yet represented here; and the implementation of this idea would cover, in theory, nearly every facet of human knowledge. There are several upper ontology projects working toward this, as I mentioned, but it appears to me that they face far more political challenges than they do technological challenges. I think that's not a good place to be.

But perhaps that can be solved. And if so, gaps will be filled. Perhaps trying to draw it all out, as at LinkedData.org, can be a red herring. The site reminds me of a printed map of internet nodes that I saw on the wall of the computing center when I was in college, in about 1990. Essentially every internet node in the US was on the map. It was big, and filled a wall, but now such a map would be impossible. Perhaps in the not-too-distant future it will be unwieldy and impractical to compile a two-dimensional display of all the linked data on the internet; as we move toward that point we'll be much closer to feeling like we have something approaching a handle on human knowledge.

I see plenty of challenges associated with these ideas. I've already alluded to some. I'd like to draw out a few others I see, and propose some potential responses. I want to mention three specific areas, and highlight two, though more certainly exist. These are Interface, Data, and Payments issues.

To start off, I'd like to look at interface issues. These aren't as much problems, as they are issues to address and clarify. Instead of having dozens or hundreds of different databases to choose from, how about just one? And, again, I'm not talking about a discovery service like those I've previously mentioned – I'm thinking of something that's a step or two beyond that. You have one interface; in fact, that's what you subscribe to – the interface, and the underlying relationships that make the content available. As I initially put this idea together, I said to myself that there wouldn't be any 'exclusives'; any vendor could access any data store. But, if access to a specific data store is controlled and limited – as it must be, if these are subscription databases and monetary and licensing issues are involved – then it is certainly possible that one primary vendor might be able to ink an exclusive deal with a secondary vendor. On the other hand, one big benefit of this project is that databases that are freely available, and use linked data, would be much more accessible than is currently the case. Using linked data would make it so much easier for the primary vendor to incorporate that data – because they

actually incorporate the taxonomy, or data structure, rather than the data itself, until just when the content is needed.

Data issues are clearly related. I've touched on this quite a bit already, and I think this is probably the biggest challenge. As I mentioned previously, there must be some form of standardization of terminologies, at the "upper ontology" level, so that systems can all link together. In the June 2010 *Against the Grain*, John Dove from Credo Reference, wrote of an initial approach at incorporating semantic web linking into reference tools. He concluded his article by stating, "I believe that if a simple set of metadata classes can be endorsed by the online reference publishing world, it will enable a whole new class of user experiences that leverage the vast quantities of reference content which is already online."¹ I think this would be a great start – but when you're dealing with reference, you're dealing with nearly the entire world of knowledge, and so creating classes that would start to work across this enormous data set is a huge challenge. If you look at XBRL, the eXtensible Business Reporting Language, which is "a language for the electronic communication of business and financial data", you find thousands and thousands of defined elements in this language alone.² There are numerous other markup languages: Chemical Markup Language; several competing languages for mono- and bi-lingual dictionaries, including dicML and XDXF; Geography Markup Language; Music Markup Language (and also Music Encoding Initiative and MusicXML); Open Scripture Information Standard; etc., etc.

Interoperability doesn't matter, though, if there's nothing to link to. How do we get as much content properly tagged as possible? What about all the old stuff that would need to be processed retrospectively? Crowdsourcing might work here, if it were easy to define the specifics of how one correctly tags content. Old Weather.org is a neat, and very new, crowdsourcing project; it is trying to transcribe data from handwritten British World War One naval logs. What's particularly notable about it, is that the people managing it have created very specific and also very helpful tools to simplify the process of transcribing this data, and collecting the most important parts of it. This is one of a variety of crowdsourcing projects put together by a specific organization called Zoo Universe; they're working on several other crowdsourcing projects, too, most funded in part by JISC.

In addition to the challenge of linking taxonomies together, there's another challenge regarding the actual accuracy of the data itself. In *Pull: The Power of the Semantic Web to Transform Your Business*, author David Siegal presents a strong argument for applying the semantic web to a wide range of applications, including the library and

¹ John Dove. "Are We There Yet, or Any Time Soon?", *Against the Grain* June 2010 (Vol. 22, No. 3): 36.

² Definition from <http://xbrl.org/whatisxbrl/>; accessed 26 Oct. 2010. Taxonomies for use in the US [XBRL has different taxonomies for different national financial reporting systems from around the world] are available at <http://www.xbrl.org/FRTApproved/>; accessed 26 Oct. 2010.

book publishing industries of the future. Though his discussion exposes significant knowledge gaps regarding libraries and bibliographic control, he does make a good point about the importance of accurate data. He writes, with reference not to printed matter but to any field trying to benefit from the application of semantic data:

Once the systems are in place, you need to focus on the three keys to success in managing a master data supply chain:

Accuracy of data

Accuracy of data

Accuracy of data³

Strangely, Siegel's solution for books is to let Google do it; the section basically consists of Siegel quoting about a dozen paragraphs from a Google Book Search product manager, who says there's no need for standards at all. There may be some connection between that and Siegel's note that LibraryThing participants make over 2000 changes to bibliographic data per day. I actually think there's a better solution to locating and incorporating ebook data, which I will address later.

There's another big challenge, I think, around financial issues. And it's not one of "how do we make this work?" Instead, it's more of an ethical issue: "should we make it work this way?" For example, how does a college manage its costs when it doesn't know what they will be? An institution would much prefer knowing its costs up front, and I've gotten feedback that that lack of control very well could prevent this idea from being accepted by libraries. But libraries are quickly adopting this paradigm, as we see from the rise of patron-driven acquisition models for ebooks. This is quite similar. It's a patron-driven acquisition model for reference information. And what is inter-library lending, but patron-driven resource acquisition? If libraries have been managing ILL costs for years, and are rapidly adding patron-driven acquisition of ebooks, why could they not manage a model like the one I'm describing?

Another option would be that one could have some sort of 'throttling down' mechanism, by which various expensive databases are not searched if the library is coming close to its monthly expense quota. And, perhaps, a library could indicate that results from the most expensive databases (per result, that is) are not ever included. A result, obviously, would be that the search done at Ithaca College or at TC3, especially, which has an even smaller budget, would not return the same results as the search done at Cornell. That's clearly a problem, since the initial goal I described was that the results would be the same on each campus. And if there is a 'throttling down' of results as the institution

³ David Siegel, *Pull: The Power of the Semantic Web to Transform your Business*. Portfolio, 2009. Location 523 in Kindle edition.

limits how much is being spent, then one would presumably see far more results for a specific search at the start of the month than at the end, with the same search. And that's probably not a result that anyone would like to see.

But what if, on the other hand, this model actually made it possible for smaller institutions to subscribe directly to fewer databases, but get more results when searches are done in the primary database? The net result in the end might make it worthwhile.

Now, understand: I'm not building this system, at least not at the moment. I have another company to run. And I haven't done any research to determine that it really would work. And some of you very well may have excellent reasons why it's a mistake all around. I'd like to hear them, and see if I can address them. But I like the overall goal, of breaking down the divisions between databases. Why should they be in separate containers? If we could use linked data to bring an unlimited number of databases together in a single interface, would that not be a far superior experience for our users?

I believe this model also has valid applications in the ebook area, as well. Ebooks are all the rage today, as they have been for quite a few years. Perhaps that's really the case now; I'll wait and see. Like all of you, I received numerous emails from ebook vendors about all the programs they're putting on at this conference, and how we should all be meeting with them at the vendor showcase to see everything new they have to offer. All the talk about patron-initiated acquisition of books certainly resonated with me as the ideas I've been putting together for today coalesced.

But one of the biggest challenges with ebooks, as with physical books, is their length, and the relative dearth of information to describe them, in comparison with journal articles. A journal article might have a paragraph-long abstract in a database; or, commonly, the entire content of the article is available in that database. A book, however, has just a few terms describing it in the catalog; its best bet is if it has a long and very descriptive subtitle. Yes, many records today do have tables of contents in them, but even then, each chapter is often about the same length as a journal article, and it still only has the description of the chapter title and its author, while the journal article has the paragraph-long abstract, or full-text.

But what a book does have, which a journal article does not, is an index. Since I imagine that most indexes these days have an anchor in the text which creates the page number as an afterthought during the process of laying out the physical pages of the book, we've already got the ability to identify the location of the discussion of people, concepts, locations, ships, and much more, within a given ebook. So, when someone wants more information about a particular subject, one way to incorporate content from monographic resources would be to provide the full text of a relevant

section of a book, as defined by its index, from the collection of books available electronically.

If this were effective for recently published books, then it would seem that it could be a way of incorporating content from older books – perhaps including all those out-of-copyright books in Google Books – as well. Simply scan in the index of the book, process it, and you have a rough method for creating linked data to and from a massive collection of still-useful published data.

The quality of linking will be directly related to the quality of an index. If a monograph has no index, even if it has the full text available, then its results won't be incorporated into this results set. Note the difference here between a morass of undefined full text, and careful definitions of the locations of relevant data (via the index). If a monograph's index is poor, then its data will similarly not work well.

Micropayments would be as relevant here as anywhere else, and they also provide particular opportunities for those who manage electronic resources to manage the related payment information, as well. If the library already owns an electronic copy of the book, then they do not need to pay anything extra to present that information to the patron. If they do not own the book, then the library would pay some level of micropayment, as they'd done for journal articles, subject encyclopedia entries, statistical data, and other electronic resources. The obvious extension to this, and the connection with many programs during the main conference, is that after a certain number of individuals had looked at a certain amount of the book (almost certainly determined on a monetary scale, rather than a "percentage of content viewed" scale) then the entire ebook is acquired by the library, and they no longer pay to access it.

And so, in conclusion, I'd like to take a step back and highlight the points that I think are most important. The overall goal is to make more information from more disparate sources more available to more patrons. Libraries have access to incredible resources, and we need to continue seeking ways to better harness and present this wealth of resources – the "fidelity," or quality, of their experience in the library – to patrons. By using just a single interface, we simplify – or, enhance the "convenience" – for patrons.

In order to make this happen, and to manage such an enormous (and always growing) quantity of so many different and disparate piles of data, I imagine applying linked data and semantic web technologies to this mass of data, to improve patrons' experiences.

Serious challenges, however, do remain. Two of the largest include figuring out how to combine numerous unrelated data sets under a single vocabulary, and ensuring that the data is represented accurately. In addition to such discovery-related issues, many access-related issues also need to be resolved, primarily around pricing, and how data could be shared and made available to as many institutions as possible.

Perhaps it is just a pipe dream, but I think it would be a big step forward in fidelity and convenience for all our library patrons, and I would welcome your thoughts on how to make it a reality.