Hi, my name is Wally Grotophorst. I’m the AUL for Digital Programs and Systems here at George Mason University.

I put this presentation together last month for our library’s administrative council. The goal was to provide staff with a “strategic update” on discovery systems—answering the question that perhaps some of you have been getting...

“So...what are we doing about a discovery layer?”

I’ve pulled a few slides out of that presentation—some that were very Mason-specific—and added a few in spots where I think it might be useful to a wider audience.
Here’s our roadmap. I’ll be dropping that last bullet point since it’s not terribly relevant to this particular venue...and I’ll apologize ahead of time if it seems the presentation ends somewhat abruptly.
This is a series of slides that I hope give you a sense of the problem we're trying to solve. I won't say these are exact, overheard quotes but if you've ever worked public services I think you’ll agree
“Oh, you should check our catalog...”
“Have you tried our e-journal finder?”
This is a series of slides that I hope give you a sense of the problem we're trying to solve. If you’ve ever worked in public services, you’ve probably said something like this at one time or another...
We get that e-journal from three different sources. I like this one...
who said using a library card was easy?
But still, researchers seem to understand and appreciate the fact that the library is still the #1 place for research ...although to the persistent researcher, it must also seem we’re #1 in ...silos
SPIRES OF EXCELLENCE!

Of course, if pressed, we’ll admit we prefer to think of them in other terms...
the catalog (OPAC)  
e-journal indexes  
licensed databases  
local e-content  
subject guides  
research portals  
campus websites  
e-reserves  
...elseweb

But whatever you call them, there is still an awfully large number of discrete places that a user needs to search to begin to feel confident that “everything possible” has been found.
Can’t we just search everything at once and not worry about where the answers are coming from?
To “search everything” we have to travel through many sources...
Maybe not as complicated as it must have been to manage filing of these cards back in the day, but difficult.

Complicated? Let’s look at our legacy information system: the catalog

It sits at the center of the librarian’s universe—as it has for many years—reflecting our history as “masters of inventory.”

Before the digital age, the catalog really did serve as an index to “what’s in the collection.” In many ways, it was a lot cleaner back then.

Today, we’ve pushed the boundaries of what “in the catalog” means. As we’ve added records for e-books or websites or e-documents or digital records, “in the catalog” has come to really mean what we can access this week. Some of these e-things come and go without anyone necessarily knowing it (this week’s subscription record load overwrites and modifies automagically).

Nevertheless, the catalog rarely gets beyond the container-level information. In today’s “digitally-enhanced” library, the catalog is an inadequate solution ... and on some level, the impulse to “get everything in the catalog” is actually making things worse.

With a cumbersome and inefficient search function, adding all sorts of disparate content on such a large scale just intensifies our signal to noise problem.
Let’s look briefly at Mason’s experience as we’ve tried improving our catalog...
Recognizing the limitations of our Voyager OPAC’s interface, we set a high priority on developing a next-gen interface...one that might introduce basic 21st century features like faceting, some social networking options, book reviews, etc.

We began by working on the open-source VuFind project...then via what turned out to be a somewhat hasty process, we decided to join our local consortium in the purchase of AquaBrowser from Serials Solutions. That decision has been, at best, a mixed blessing.

We have a functioning, 21st century interface to our OPAC with AquaBrowser (along with that @#)*&! FLASH word-cloud widget) but we’re also the victim of what I guess you’d have to call a blend of vendor abandonment and upsell.

Not six months after AquaBrowser came online, we got this word from SerialsSolutions...[next slide]
“We will continue to support libraries on AquaBrowser v2 at least through 2012. We recognize that you are heavily invested in your current AquaBrowser and, for 2010, you will probably want to keep using the AquaBrowser v2 that you installed and customized...

We'd be happy to talk to you about subscribing to the new product in 2011, or sooner if you desire.”

-- SerialsSolutions, 2010

Have not been able to find the name of any site using Aquabrowser V3 (SaaS).
We don’t have to start over but it seems we have to keep moving

So, with the future of Aquabrowser in doubt, we have an interim fix and a nice mobile catalog (thanks to the more future-oriented design of Aquabrowser, we use it as the target for our mobile catalog).

http://mcat.gmu.edu
Strategy?

Continue with Aquabrowser as interim solution.

Continue local development of VuFind

Realize that we’ll likely move to a new platform in 12-18 months

So, what do I recommend in terms of a discovery solution for our OPAC? More on the “continue with VuFind” in a moment...

<yeah, the screen is from Civ V>
Beyond the OPAC

Let’s expand our discussion on Discovery
Every discovery solution relies either exclusively or more exhaustively on one of these two design principles...
Just in Case or Just in Time
Search service crawls and indexes content before the search begins... just in case it might prove useful

You don’t search the actual content, you search the cached metadata which contains links back to content
Here's a diagram of a JIC system. This is actually a diagram from the paper that Sergey Brin and Lawrence Page wrote when introducing google.stanford.com to the world.
An interesting read if you like to geek out on search engines and internet history. Google is the canonical JIC search system. http://infolab.stanford.edu/~backrub/google.html
Search is real-time. You query the search service and it queries "live" content sources.

JIT or "Federated Searching" is slower and limited by latency introduced by queried sources.

JIT model works best for "deep web" content.

Just In Time is a very different approach. This is federated searching.

So, JIC systems are fast but you sacrifice currency (reliability) of information. You can retrieve an item only if it was collected and indexed prior to your query. If it just appeared on the web, it’s invisible to you. By contrast, with a JIT system you sacrifice speed to improve reliability (currency) of information.
The three vendor logos at the top of the screen illustrate companies with JIT solutions.
Here are some examples of JIC, JIT and hybrid discovery systems. The *'s mean the systems combine features of JIC/JIT or locally developed code in some way (typically by integrating some sort of federated search).
So, having skimmed over JIT and JIC systems, let’s look more closely at Summon. Summon is adamantly 100% JIC -- everything has been collected and indexed before you begin a search.
ProQuest leverages the content it controls (e.g., Chadwyck-Healy, Cambridge Scientific Abstracts, ProQuest, UMI Dissertations, article-level metadata from SerialsSolutions, etc.) as well as content contributed by some of its competitors and other sources to build a single, unified index that sits behind a “single search box.”

Competitors? Sure. ProQuest pitches Summon™ to competing content providers as a way to boost renewal rates and usage, promising to drive customers their way with outbound links once they’ve given ProQuest indexing access to their content and/or metadata. As they make quite clear to potential partners, “Your full text content will never be displayed in the Summon™ service”.

http://www.serialssolutions.com/summon-become-participant/
Why we like it...

There’s not so much to explain

The library has access to all of the content in the system, minimizing user frustration

It is fast, faceted and looks great

Don’t use Google™ for your research, use our ‘library google’ instead

It’s not at all hard to see why many public services librarians find Summon compelling:
But there are a few things you need to think about...

If you’re curious, this graphic comes from a german book on electricity. I’ll refer my German speaking friends here: http://www.tmw.ac.at/default.asp?id=297&al=Deutsch
Searchers will miss whatever ProQuest leaves out of the Summon\textsuperscript{TM} database.

I don’t know why but it seems the more I think about what “discovery layer” has come to mean, the more often I think aquaculture offers an appropriate metaphor? Our “discovery” resource free-floating on a sea of content.

Beyond not really knowing what’s inside the nets this week, it’s also not clear whether Summon has indexed just metadata for a source or the full-text. Not only does that affect retrieval but it also has a huge impact on relevance ranking algorithms. With Summon, that’s all taking place below the surface and it is invisible to us.
By design, the system can never be absolutely current.

Content has to be delivered to ProQuest, then indexed before it appears in the Summon knowledge base (index).
With a product like Summon™, we’re saying, “if we don’t have it, you probably don’t need it”

This is something new for libraries and I’d argue it threatens to upend at least two generations worth of practice.

Some might say we’re not saying that at all...as in, “we have librarians who can explain the limitations of the service to users” and that’s certainly true. But if you promote something like Summon™ as the “single search box” how is it that you aren’t at least subliminally sending that message?

Recognizing some of these issues, there are few ways that ProQuest has started dealing with them, and I’ll get to a few of those in a moment...
Use Summon™

• As a starting place for your research
• If you need to quickly retrieve high quality items on a particular topic (e.g., "I need three good sources on ______.")
• If you are interested in interdisciplinary and multidisciplinary research (e.g., "I'm looking for information about how global warming is affecting the health of populations in developing countries.")
• If you know the citation of an item and just want to get that article (e.g., "My professor asked me to get all the articles that are listed in this bibliography.")

Does implementing a discovery layer mean the end of having to help users figure out how to use our systems? Not really, it will just change to things we have to teach them.

AU offers its users these reasons for using Summon™

http://www.american.edu/library/research/searchbox.cfm
Avoid Summon™

• If you need to retrieve statistical/financial data, do not use Summon™

• If you need to perform in-depth searches that require extensive use of a database's controlled vocabulary, do not use Summon™.

• If you need to retrieve results from art image databases and not just articles about art.

• If you’re searching for a current event, remember that Summon™ is not a current affairs database

As well as these cautions...
If you worry most about helping the user who asks, “find me something useful” then Summon™ is a winner.

If your job depends on satisfying the user who asks, “find me everything” or “is this absolutely current?” then Summon™ can be a distraction.

My sense of how something like Summon shakes out...

The “I need three articles on climate change” user will love Summon. The librarian who just points that user to the “single searchbox” will love it. It is a very good product for “find me something useful” crowd.

Not so great if you’re an art student looking for images or a researcher trying to keep up with a fast-moving, relatively narrow field. In late June, SS and ArtStor signed an agreement which will presumably improve Summon’s support for that art student.

At some level, it seems the utility and value of something like Summon™ is inversely proportional to the sophistication and information needs of the user.
The corporate parent of Summon™ is in the business of **selling** content. Do we just **trust** that ProQuest will never:

...bump up its own holdings higher in the retrieval sets (to encourage library renewal of subscriptions)?

...find subtle ways to penalize content from competitors?

But here’s where I really think we have a really huge problem. If you want a label for it, try “content neutrality”

Buying your search engine from the company that also sells you the content does more than just disintermediate librarians.

This slide refers to Summon™ and ProQuest but the same is true of EBSCO and their EDS discovery product (e.g., did you notice they just purchased H. W. Wilson ...another step in the long march to their version of e-content hegemony).

Of course, as we’ll see in the next few slides, from time to time even the most enthusiastic booster of ‘discovery’ layers has to notice that it’s not always “all about discovering”
In 2009, ExLibris was so happy that EBSCO was allowing them to put the EBSCOhost metadata into the Primo index that they issued this press release.

Ex Libris Partners with EBSCO Publishing to Offer Primo Users Online Access to Authoritative Electronic Journals via Primo Central

Primo Central will index EBSCO content centrally and make it available to Primo users with other library and institutional collections.

Chicago, IL—July 9, 2009. Ex Libris® and EBSCO Publishing (EBSCO) have signed an agreement that enables EBSCOhost® subscribers that run the Primo® discovery and delivery solution to offer their users seamless access to the authoritative EBSCOhost electronic content via a new Primo component, Primo Central.

This flexible Primo technology will be used by Ex Libris to index the EBSCO content centrally and make it available for users’ searches along with all other library and institutional collections. Primo will display the search results in a single relevance-ranked list.

EBSCO’s trustworthy content, which includes some of the most important databases for scholarly researchers, will be more visible to library patrons through Primo. Ex Libris and EBSCO will ensure that libraries that use both Primo and EBSCOhost services are able to provide their users with seamless online access to the full-text information that they want, whether it is discovered through Primo, EBSCOhost Integrated Search, or EBSCO Discovery Service™.
As you may know, for the past eighteen months, we have been indexing in Primo Central a number of the EBSCO databases. EBSCO has now changed their strategy and will no longer permit third-party discovery services to load and index their content. Therefore, starting 1st January 2011 we will cease hosting of the EBSCO content in the Primo Central Index. EBSCO will, however, permit our use of a specialized API to search the EBSCO content ‘just-in-time’.

Since our initial agreement with EBSCO in June 2009, we have made significant progress in working directly with many publishers and other aggregators to dramatically increase the content in the Primo Central Index. In addition we recently reached agreement with Gale whereby their databases in Primo Central will now be available to all, regardless of subscription. Since there is a considerable overlap between some of Gale’s and EBSCO’s collections, EBSCO subscribers will benefit considerably from Gale’s consent to open up their data. Furthermore, Gale's move indicates the general trend of information providers of enabling their data through multiple distribution channels and we are delighted to witness this change.

Based on a recent analysis of the Primo Central content, we cover, through other channels, over 90% of the data provided by the current EBSCO content loaded in the Primo Central Index. Furthermore, of the small number of titles exclusively available from EBSCO, none of these appears on the list of the 5,000 most used journals, based on SFX logs, and only three appear on the list of the 10,000 most used journals.

We are currently finalizing the details of the new arrangement with EBSCO for ‘just-in-time’ search and will update you as we progress on this. However, we believe that EBSCO’s decision to withdraw their content from the Primo Central Index does not best serve your user’s interests. We therefore strongly encourage you to add your voices directly to those of the ELUNA and IGELU steering committees in requesting that EBSCO reverse their decision and enable their data for indexing.

by 2011, EBSCO (now selling their own discovery platform) decides to deny ExLibris access to their metadata.
As you may know, for the past eighteen months, we have been indexing in Primo Central a number of the EBSCO databases. **EBSCO has now changed their strategy and will no longer permit third-party discovery services to load and index their content.** Therefore, starting 1st January 2011 we will cease hosting of the EBSCO content in the Primo Central Index. EBSCO will, however, permit our use of a specialized API to search the EBSCO content ‘just-in-time’.

Since our initial agreement with EBSCO in June 2009, we have made significant progress in working directly with many publishers and other aggregators to dramatically increase the content in the Primo Central Index. In addition we recently reached agreement with Gale whereby their databases in Primo Central will now be available to all, regardless of subscription. Since there is a considerable overlap between some of Gale’s and EBSCO’s collections, EBSCO subscribers will benefit considerably from Gale’s consent to open up their data. Furthermore, Gale’s move indicates the general trend of information providers of enabling their data through multiple distribution channels and we are delighted to witness this change.

Based on a recent analysis of the Primo Central content, we cover, through other channels, over 90% of the data provided by the current EBSCO content loaded in the Primo Central Index. Furthermore, of the small number of titles exclusively available from EBSCO, none of these appears on the list of the 5,000 most used journals, based on SFX logs, and only three appear on the list of the 10,000 most used journals.

We are currently finalizing the details of the new arrangement with EBSCO for ‘just-in-time’ search and will update you as we progress on this. However, we believe that EBSCO’s decision to withdraw their content from the Primo Central Index does not best serve your user’s interests. We therefore strongly encourage you to add your voices directly to those of the ELUNA and IGELU steering committees in requesting that EBSCO reverse their decision and enable their data for indexing.

I’m still puzzling over just what “third party discovery service” means in this context. I assume it means quite specifically someone like ExLibris—a discovery layer competitor that’s not also engaged in publishing.
Summon™ now offers one or more suggestions for specialized sources.

Now, getting back to the ways in which SerialsSolutions is working to improve Summon.

Responding to some concern over the way “discovery layer” can actually obscure information, Summon™ now offers outbound links to more specialized sources.

As far as I’ve been able to determine, it seems that Summon’s outbound links are limited to various EBSCOhost databases...and I hope that I’ve just missed the fact that they offer other resources.
Another option to extend Summon: “Add results beyond your library’s collection.”

[click to add red circle over facets and highlight results.]
Notice we have 135,391 results before checking the “add results beyond your library’s collection” box.

And 62,857 journal articles in that group of 135,000 objects.

Does this mean Summon is about to scour the web for more content? Well, not exactly. Instead, you’re about to be shown the rest of the Summon index (whether your library is authorized to follow links to the content or not).
Two points to notice:

1) this new content isn’t a single facet so you can’t easily see what’s just been added

2) the 66% jump from 62,000 journal articles to 103,000+ suggests to me that a lot of content the NCSU libraries might not be all that interested in, has dramatically increased the size of the results set.

In a library with a weaker collection, a bump of this scale might make perfect sense...but is an ARL library and I’m guessing there isn’t a way to increase relevant articles by 66%. Is Summon “spamming” the results set with this function? Hard to say.

Without a way to isolate or drill down into the “beyond your library’s collection” set, there’s just no easy way to know the answer to these questions.
I think the idea we see taking shape at Villanova’s goes a long way toward solving some of the serious issues I see in these discovery products. And sites like NC State with their Endeca OPAC integration follow a similar path -- as presumably will UVa and their integration of Blacklight and Primo.

Villanova is using VuFind to provide a faceted search interface for their Voyager system and then using the Summon API to provide other e-content in a separate pane. One search--two sets of results.

The upside to this method? Items from the catalog--presumably better reflecting the research focus of the institution than the e-content in Summon--aren’t buried (or intermingled) with possibly irrelevant hits from the e-content search of Summon. Both Serials Solutions (Summon) and ExLibris (Primo) offer this sort of API access to their content and I suspect going forward building hybrid systems that take advantage of the Summon or Primo API may become more popular.

I wish I had more details on whether these discovery vendors will unbundle their products, for example, allowing a library to purchase API-only access to their content database.
Unlike some vendors who cloak their API documentation behind an NDA, SerialsSolutions offers an example I wish more vendors would follow: documenting and assisting with access to the API for their content index. This openness has already fostered a number of advances and I think it’s critical for sites that might want to incorporate the strengths of the Summon database but develop locally to mitigate the limitations.
First, here’s an article that Michael Klein from Oregon State published in Code{4}lib journal, detailing how they’ve used local programming to improve their Summon implementation.

The University of Michigan, a Summon customer, has released a Drupal module for accessing the Summon API.

This last screengrab shows an example of the PHP code posted by SerialsSolutions on their API documentation site.
I’ll stop here as the rest of the original presentation was very Mason specific.

Wally Grotophorst  
Associate University Librarian  
Digital Programs and Systems  
University Libraries  
George Mason University  
Fairfax, Virginia 22030

e-mail: wallyg@gmu.edu  
twitter: @grotophorst