How to Read 100 Million Publications: VIVO and Comprehensive Open Publication Databases

John Mark Ockerbloom
Good morning. I’m so glad to be a part of this session, and get to hear about the work that people at Notre Dame and Florida have been doing with their partners. It’s clear they’ve put a lot of thought not only into how they’re going to compile and use data in their own institutional services, but in how they’re going to work with outside partners like SHARE and CHORUS to make the data in those services more broadly usable, and cover a wider range of information.

That’s tapping into the real value of linked data, the kinds of interactions that Dr. Soren Auer showed in his talk this morning. It’s not to say that linked data like RDF and SPARQL aren’t important, but they’re just a means to an end— they’re ways of enabling us to share, reuse, and enhance information that’s of interest to large communities.
That’s why when I look at our VIVO instance at Penn, which is one of the larger ones out there, I’m impressed with everything our faculty is doing and publishing—our VIVO has over a quarter million publications and counting. But I’m also thinking it’s really too small.

Or at the very least, it’s too small to take full advantage of the data we’ve compiled. It’s a good size for showcasing the work of our medical school, and our health sciences faculty. But it’s also more isolated than I’d like from the rest of the research community. Sure, it’s great that our faculty are collaborating with folks from Cornell, and Florida, and Colorado, and other places. But in my VIVO, I can only easily see those collaborations from Penn’s perspective.

Those collaborations are also represented in those other institutions’ VIVOs as well, but I can’t easily see or link to the information they’ve collected. I can easily see who one of my faculty members and one of Cornell’s faculty members both work with. But it’s harder to see who that Cornell faculty member works with when she’s not working with Penn faculty.

And even on the direct collaborations, we’re spending a lot of effort trying to pull together the same information in different locations. The venues for our co-authored papers should be the same in both Penn’s and Cornell’s VIVO, but we each have to get this information separately, and each of us maintains our own triples for the same journals and conferences. There might be also open access copies of those papers in one our institution’s repositories, but not the other. It’s be really nice to record information like that in both of our VIVOs, but it’s hard to share that sort of data.

In the VIVO community we do put a lot of emphasis in other kinds of sharing and collaboration. We share the software we develop. We collaborate in designing ontologies, and we share those as well. But when it comes to the actual data, it’s largely each institution for itself. Now many of us work with commercial partners to wrangle our data, and that’s valuable help that we appreciate. But I think there’s also more we can do as a community to grow our VIVOs more effectively with shared data.
I think we need to be thinking bigger about distributing and aggregating the data in our VIVO instances. We know that our researchers like to work with big aggregations, for instance. When they search for new research in their field, they generally don’t go to their local VIVO site. They go to Google Scholar, or Web of Science, or PubMed, or just ordinary Google. And they have good reason for that— they get access to a lot of publications at once. Now, if we’ve done good search optimization on our sites— which is something that Anirvan Chatterjee can tell you about this afternoon— maybe they get led back to our institutional sites. But often these big aggregators won’t take people there.

We also have some big questions ourselves about our scholars’ work that go beyond questions of finding and sharing individual articles.
We also have big questions

- How much “library-interest” scholarly communication is being reliably preserved?
  - And how can we improve on that?
- How much of it is open access?
  - Likewise, how can we improve that?
- What text/data patterns emerge across a wide research question/field of research?
  - Mining needs representative corpuses/samples
- How are researchers collaborating outside our institution, or across an entire field?
  - Classic VIVO question, scoped beyond one org’s VIVO

As librarians, we want to make sure we can keep providing access to the research that our scholars produce and use far into the future. And one of the scary things we ask ourselves, and that big sites can’t easily tell us, is how much of the increasingly digital content that we collect is being securely preserved by third parties. Will we still be able to get to the research that was published in journals if their publishers go under, or we have to drop subscriptions, or some other calamity happens? That may depend on whether it’s been backed up by LOCKSS, CLOCKSS, Portico, or some other preservation system. But do we really know how much of our content is securely backed up that way?

We also want to make the most of the content we take responsibility for. Many libraries would like to make as much of it as possible legally readable by the entire world— but it’s often hard to tell what’s openly accessible in some version, be it preprint, postprint, or version of record— or what *could* be made openly accessible if we had more information about publisher’s contracts and copyrights.

Also, our researchers are increasingly wanting to mine publications, having machines scan a large body of publications at once, to find patterns in the text or the data. One of the requirements for reliable mining is that you need to have a representative sample to mine. A digitized corpus that’s largely full of material from medical libraries is going to have rather different textual characteristics than a corpus full of materials from liberal arts colleges. If we have a limited collection of material from the scholarly record to analyze, we need to understand how that collection relates to the scholarly record as a whole.

Finally, when we’re studying collaboration among researchers, we might want to understand how researchers work together across an entire multi-institutional community, not just when they’re working with one our own institution’s researchers. This is one of the classic VIVO questions, scaled past individual VIVO instances.

A common theme with all of these questions is that a widely scoped comprehensive database about scholarly publications can help you answer all of these questions more authoritatively than our existing systems allow.
So how big would a truly comprehensive record of scholarly work be? As you might have guessed from the title of my talk, it’s around 100 million publications. It’s hard to be very precise about this, but there’s a 2014 paper by Khabsa and Giles in PLOS One where they estimate that “at least 114 million English-language scholarly documents are accessible on the web, of which Google Scholar has nearly 100 million.”

We see similar numbers for some other large records of scholarship. CrossRef has stated they track over 80 million publication DOIs; Web of Science recently claimed more than 90 million publications, and Microsoft says they have over 80 million papers indexed in their Academic Search.

I’m sure we could get bigger sizes if we went further into the past, or more into grey literature, or dug more deeply into foreign languages. But 100 million would seem to be about the right order or magnitude to consider.

Now, to be clear, I’m not saying that a shared database of publications is useless until you reach this size. There are a lot of things you can do with a more limited database, and I’ll show you some examples in a bit. But if we’re considering supporting a shared knowledge base of publication information, it’s worth planning for an infrastructure that’s capable of reaching this size.
From a purely technical standpoint, 100 million data objects isn’t too big to manage.

I mean, it’s way bigger than any VIVO I know of. It’s bigger in fact than all of the VIVOs I know of put together. And I don’t think I’d want to go and load 100 million publications into my VIVO instance at home. I admit I haven’t tried 1.9 yet, but the earlier VIVO versions I have tried haven’t really been designed for that scale. A quarter-million publications, fine; 100 million, probably not.

But there are certainly other technologies that can handle that size. Pedro Szekely’s giving a talk later today about semantic web technology that he says scales well up to billions of semantic triples— that’s the sort of scale that you’d have for 100 million publications. And basic metadata on 100 million publications also fits comfortably on commodity disk storage. Even if you average 10 kilobytes of metadata per publication— which is a lot more than you generally need for the basics— 100 million metadata records fit into 1 terabyte.

We may even be able to pass it around the metadata with ordinary Internet tools. Microsoft publishes dump files with essential metadata on all of the more than 80 million publications in its Academic Search database. I’ve downloaded their largest dump recently – it’s a little less than 30 gigabytes. With wget or a BitTorrent client I can fire up a download and have it all on my workstation at Penn within a few hours.

Today, the main challenges with this size of data is not so much storing it, or computing with it, or distributing it. It’s getting the data together in the first place, and ensuring that it’s workable enough for what you want to do with it.
How do we get data that big?

- **Collect it yourself?**
  - Can get it all from users, if you’re where “everyone” goes
    - E.g. Facebook, Github, NIH
    - But individual institutions, even VIVO community as whole? No.

- **Aggregate it opportunistically**
  - SHARE, ORCID, BASE doing this
  - Google/Google Scholar do this at larger scale
    - Again, works if you have very high capacity, otherwise no.

- **Aggregate it systematically**
  - This can be more suitable, feasible for specific issues
  - ...but prepare for messy data (unavoidable at this scale)

- **(In practice, often a mix of strategies)**

One of the liberating things that happens when you plan for 100 million data objects is that you can instantly discard some ways you might think of managing the data that are clearly too labor-intensive. For instance, you know that there’s no way you’re going to collect 100 million citations yourself– the scale’s just too big. Maybe you can collect data from your users if you’re a place where everyone goes; like Facebook with its billion-plus users, or Github with its 38 million developer repositories, or NIH where every American medical researcher would like to get a grant, and they have to supply you a biosketch to get one.

But we don’t occupy central roles like that, either individually or as a community. Maybe we think “Oh, we can use volunteer crowd-sourcing to get data”. And you can for small amounts of data.
For instance, in genre fiction there are some big open indexes run by amateurs that now track over a million popular magazine stories and articles from the 19th century to the present day. They track authors, versions, and venues of pretty much any piece of magazine fiction that you’re likely to find in a library – and many pieces that you won’t. If we had scholars or librarians that did this thorough a job with research journals, and provided easy ways of acquiring and using that data, I’d be really happy.
But even the biggest crowdsourcing projects have their limits. Wikipedia is one of the most visited sites on the Internet. Tens of thousands of people edit it regularly, and it’s been encouraging editors to add citations to its articles for many years. So how many have they collected? I ran a script on their most recent monthly dump and it looks like the English Wikipedia has about 13 million instances of their standard citation template. That’s an impressive number, but it means that a very big crowd has still produced well under 100 million cites.

(It’s not like they’ve run out of places to add citations—there are also over 300,000 articles in English Wikipedia that currently include a “citation needed” notice.)
In practice, if you need a *lot* of publication data, you’re going to need to get most of it from other sources that have collected it already. You need to aggregate it, in other words. There are two basic approaches you can take. You be opportunistic, and collect data from wherever you can easily get it, and not worry too much about data you can’t easily collect. Or you can be systematic, and try to figure out what data you most need to address whatever issue you’re most interested in, and then try to collect as much of that as you can, whether that’s easy or hard.

So if you’re interested in data on influential scholarly publications, you could start by collecting records from projects like SHARE and BASE, which have been harvesting metadata from repositories with OAI. That will get you information about a lot of articles, but it won’t necessarily represent the most influential ones– just the ones that have been deposited into public OAI-compliant repositories. If you think that the most influential publications tended to appear in high-profile journals, though, you might want to compile a list of those journals and then do whatever it takes to get metadata from them. Sometimes that will be easy, sometimes that will be hard, and you probably won’t get to 100 million publication records that way. But the data you do collect may be better suited to answering your question than a larger but opportunistic data set.

This is a bit of an artificial dichotomy– in practice a lot of knowledge bases collect both opportunistically and systematically. For instance, ORCID’s goal has been to assign a unique identifier to every currently publishing scholarly author. They started out by simply inviting people to sign up and create an ID– that’s an opportunistic way of building up an identifier set based on interest. More recently they’ve teamed up with publishers and funders to require that everyone who submits to them gets an ID– that’s a more systematic approach, since it ensures that ORCIDs will eventually be assigned to everyone who publishes in the journals they target.
So how do you decide what your aggregation strategy is going to be? Here are a few questions that can help you plan it.

First of all, what are you trying to accomplish by aggregating the data? And is someone else already doing that? If they are, you might want to just let them do their thing. For instance, don’t try to out-Google-Scholar Google Scholar. For better or for worse, Google has shown that comprehensive general search is largely a winner-take-all competition, so that’s probably not the best place to put your energy. For similar reasons, don’t try to out-Web-of-Science Web-of-Science.

But we’ve seen that there are a lot of other things we want to do with scholarly publications besides the things they do. Let’s take a preservation question we asked earlier: What proportion of some set of journals we care about is being backed up in long-term preservation repositories? To answer that, we need to know both what’s in preservation repositories, and what makes up the journals we care about. And some of that data’s already been collected...
...in particular, there’s a project Keepers Registry has a database of long-term preservation repositories and what’s in them. So maybe we can use their data.

But – can we use that data? We can look up journals on the website here, but can we extract the data in a usable form so we can easily compare it against a journal set?

Well, kind of. The Keepers Registry doesn’t have a bulk download facility that I’ve been able to find, but it does have some APIs I can use to extract information about journals I’m interested in. When I look at that data, I find that a lot of the journals aren’t preserved as full runs. Instead, typically a certain range of its volumes might be stored in various places. So I’m also going to have to find out or guess what volumes my journals have before I can find out what proportion of them are preserved. It looks like I can’t get *that* information from the Keeper’s Registry, but I might be able to get it from somewhere else.
Some basic data for systematic scholarly publication aggregation

- **A set of venues**
  - Scholarly articles: Journals, conference series, paper series
  - Broader scope might include sharing sites
  - Many of these have standard IDs! (E.g. ISSNs) But not all do.

- **Venue bundles**
  - Journal volumes + issues, conference meetings...
  - Usually no standard IDs, often standard descriptions (volume, issue, date...)

- **Publications in these venues and bundles**
  - Articles, monographs, data sets, presentation materials.
  - Some basic table-of-contents properties:
    - Title of publication
    - Authors of publications (names good, IDs better)
    - Venue/bundle in which they were published
    - Page range, section, or sequence designation (if applicable)
    - URIs for publication versions/components (both official & unofficial)
  - Standard IDs common, but far from universal: DOI, ORCID, VIAF...

A comprehensive collection of basic metadata about publications, for instance, could answer our questions about volumes of a journal. And it could also help answer a number of our other questions about the scholarly record.

Ideally, we want to collect and share basic information about journals (and other scholarly venues), about volumes (and other kinds of venue bundles), and about the individual publications in those volumes.

If we have a list of journals with standard identifiers, such as, say, ISSNs, then we can automatically match those journals up against the journals mentioned in the Keepers Registry.

And if we have a list of volumes of those journals, we can compare that set of volumes to the volumes the Keeper’s Registry knows about. Unfortunately there aren’t pre-existing standard identifiers for volumes or other journal sections, but there are standard ways to describe them, like volume number, issue number, and date. And if we’re maintaining volume information as linked data, then the URIs we use for the volumes can potentially *become* standard identifiers for them that we can use in other contexts.

For this preservation question, we don’t have to know much about the actual articles that appear in the volume. But we might want to at least know how *many* articles appear in each volume, in case we want to weight volumes by their relative size. And for other contexts, it’s good to have some basic metadata about articles, like whatever you’d say about them when citing them. It would also be good to know where various versions of the article can be found online. Standard identifiers like DOIs, ORCIDs, and VIAF IDs can help us correlate data about articles and their authors in multiple systems.

Are there databases out there that have this kind of data? Well, there’s actually a bunch of them. Crossref has records with a lot of this kind of metadata for the DOIs that publishers register with them. You can also find this kind of data in Microsoft Academic’s datasets and in PubMed, Web of Science, ORCID, individual publisher platforms— and to some extent in our various VIVO instances. Of course, it’s easier to work with some of these sources than with others.
Can you reuse the data?

- **What data sets are available?**
  - Structure? Quality?
- **How can you get it?**
  - APIs? Bulk downloads?
- **What are you allowed to do with it?**
  - Is it copyrighted? Copyrightable?
  - Does supplier impose restrictions on it?
  - Does supplier certify it as open?

One database, for instance, might severely limit what sorts of data you can automatically extract from it, what you’re allowed to do with the data, or whether you’re allowed to redistribute it. Another database might not put any restrictions on how you use their data, but it might be harder to work with because it uses unfamiliar data structures, or its data is messy. Mind you, pretty much any database with 100 million objects in it is going to include some messy data, but some kinds of mess are easier to deal than others.

At any rate, for any data source you want to aggregate you should ask: What do you get; how can you get it; and what are you allowed to do with it.

The Keeper’s Registry for instance, has a Terms of Use page that says that many of their data sets come from proprietary sources, that we can’t pass along any data that violates the requirements of those proprietary sources, and that it’s our responsibility to find out what those requirements are. So while I might I use their data to produce summaries of preservation coverage that I might share with the world, I wouldn’t attempt to add their raw data onto a open publications database until I cleared things up with the appropriate providers.
Okay, let’s say we’ve found some usable sources for aggregating data. We probably still have some work to do. We might need to clean up some of the data we got, for instance, or enhance it. We don’t need 100% pristine data in many cases. If our sources don’t have intelligible volume counts for some of the journals we’re interested in preserving, we could try to go collection some additional volume data ourselves. Or we could just automatically estimate how many volumes those journals are likely to have, maybe based on when they started publishing, and that’ll give us a coverage answer that’s accurate enough for our purposes.

I’ve talked a lot about aggregating data from a consumer’s standpoint, but if you’re going to be maintaining a large database over a long period of time, you should really be thinking in terms of collaborating with the key sources you’re aggregating from, and not simply consuming their output. You’re depending on them to give you stuff, and that probably takes some effort and resources on their part, and maybe some risk. It’s nice to make your relationship a symbiotic one. If you work with them and support them, you make it more likely they’ll keep supplying the data you need, and not try to stop you from using it. Maybe the corrections and enhancements you make to the data you collect would improve their data as well. And maybe if you tell them how you’re using their data they can find better ways of maintaining and supplying it to meet their needs. Or they can tell you how you can make better use of the data they’re already providing.
Let’s not have 15 competing data silos

The risk of not building up collaborative relationships with your data sources – or your data consumers – is that you risk building up yet another data silo. That’s not what we want. There’s a well-known XKCD cartoon that talks about how people waste time producing redundant and competing standards– we don’t want to do the same sort of thing producing redundant and competing data silos. We want to work together wherever that’s feasible.

That symbiotic principle applies to VIVO development in general too. If you find VIVO useful, I encourage you to consider ways you can help sustain VIVO development.
All right, end of plug. Let me talk about another publication data project that demonstrates useful collaboration principles.

The Global Open Knowledge Base collects and publishes information about journals and other electronic resources that libraries can acquire for their collections, in both human and machine-readable form. It includes basic information about journals including their titles, identifiers, time of publication, relationships to other journals, and their availability in various subscription packages and online platforms.

The data is as open as they can manage – it’s CC0 licensed, and it’s currently available to the public via an OAI feed, and will also be available through other mechanisms, including linked data, in the near future.

It’s a collaboration with various publishers and directories (who supply data about their offerings). A lot of them are happy to do so because this database becomes a catalog of offerings for customers to buy.

It’s also a collaboration with funders (initially Mellon and OLE, and in the future the Open Library Foundation) that support the system development

And also a collaboration with librarians at NC State and elsewhere that correct and enhance the supplied data.

Because they’re dealing with publication information at the journal level, rather than the article level, the editorial work they need to do on the data is kept manageable. Tools like OpenRefine also make it easier to apply data corrections and enhancements in bulk

Although this journal-level data is primarily intended to serve the needs of acquisitions, it’s general enough and open enough that it can serve a number of other purposes as well.
For instance, it might be useful for supporting open access services. I should mention that for a long time I’ve maintained a browsable directory of free open access books and serials online. It has millions of items in it, and the vast majority of the data in it has been automatically aggregated from sources like HathiTrust and the Library of Congress Subject Headings. Along with the aggregation make a relatively small number of local enhancements in response to my users’ requests and feedback. My manually added data makes up maybe 3% of the database at most.

Lately I’ve been adding more support for serials access. I added ISSNIs to my database so I could easily aggregate from the Directory of Open Access Journals, for instance....
...and I’m starting to contemplate ways one can use data from GOKB to bootstrap finer-grain access to journals that are only partially open. For instance, you could use GOKB’s ISSNs and platform information to find ways of getting article level metadata from a journal’s platform (or some other source that lets you look up and reuse article information by ISSN.) Some of those articles may be open access at the publisher’s site. Others might not be open there, but have a preprint that’s available in an open repository. You could potentially correlate journal article metadata with metadata in systems like SHARE or BASE, to help identify and locate other open access versions of certain journal articles.

That sort of data could not only support an open access-friendly browsing service like the one I’ve mocked up above, but also answer questions like how much of this journal’s scholarship, or its field’s, can be found in some form via open access. Maybe we could even publish open accessibility scores like the ones you see on this screen.

And if we’re not under supplier constraints, we wouldn’t keep this data to ourselves: we’d make it easy for other users to get it, maybe with APIs, or JSON-LD formats, or monthly dump files. Maybe a data mining project could use this data, for instance, to come up with a representative sample of library science articles or data that it can then retrieve and text-mine to find important trends in the field.

Now, I’m not saying that the next thing I should do, or that we should do as a group, is drop what we’re doing and build this. There are complications in implementing a system like this that I don’t have time to describe right now, and we’ve all got lots of things on our plate. And for all I know maybe SHARE or some other project I haven’t heard of is already working on even better open access dashboards. Or maybe when table of contents set like this isn’t easily obtainable, you could still calculate open access scores with more limited data, like comparing the counts of distinct open access articles that appear to come from this journal with a count or estimate of the total number of articles a journal published.

In any case, I want to show it as an example of how broad aggregation, focused enhancement, and open, machine-friendly redistribution could enable a variety of useful services and answer big questions across a wide range of publications.
I’d like to wrap up with a more VIVO-specific example.

We’ve had some interesting demos of aggregators of VIVO data, like VIVOSearch, but mostly they’ve been demos that have come and gone. Don’t get me wrong—OpenVIVO is a really

But there’s on important aggregator of data largely from VIVO that’s still in production, and that’s CTSASearch. It’s not actually sponsored by VIVO or Duraspace as far as I know, but maybe it should be. It lets us find investigators and explore collaborations more widely than any single VIVO site lets us. And it works by aggregating data from lots of external sources, mostly from VIVO sites. It also has made some data enhancements that are of interest to a CTSA audience, like identifying and mapping investigators in certain health research programs like BD2K and FORCE2016.

And it’s pulled in a fair bit of publication information—about 7 million publication records. That’s not big enough to make it rival the big publication databases out there, but it does make it big enough to be interesting for studying CTSA activity. There’s probably some redundancy in those 7 million records, since there may be co-authored paper represented in multiple VIVOS, but perhaps a publication database like the one I’ve described earlier on could help in automatically cleaning up that redundancy. And a deduped, cleaned up CTSA search can also help individual VIVO instances clean up their data as well, and establish links across instances when they have things in common like publications or people. In fact, there’s a talk later today about CTSASearch and how it’s been used to establish cross-institution Profiles links.
I hope that these examples show some ways a shared open database of publication information could be useful to those of us who run VIVO instances and repositories -- and perhaps also show some ways that the data in VIVO instances could support such a shared database. I’ve also shown how compiling a database of 100 million publications is not unthinkable, if we can make the most of aggregation, focused data enhancement, open sharing of data, and support of a robust data exchange network.

Whether or not we compile a shared database that big, *thinking* that big can help us design research support systems that can scale to that size when appropriate. And it can help us envision and eventually implement applications that can make the most of a comprehensive open publication database.

If that resonates with anyone in the room, this is a really good opportunity to discuss how the VIVO and Duraspace community could work together, and how we could work with publishers and data aggregators, many of whom are represented at this conference. Whether you want to discuss priorities, brainstorm more ideas, talk about the practical hurdles to overcome in compiling open metadata, or propose adopting and supporting some of the ongoing projects I discussed like GOKB or CTSAsSearch, I’d love to hear what you have to say.

Thank you.