August 16, 2013

How to Read 200,000 Publications: VIVO and the intelligent evaluation of scholarship

John Mark Ockerbloom, University of Pennsylvania
Anne Seymour, University of Pennsylvania

Available at: https://works.bepress.com/john_mark_ockerbloom/15/
Thank you for coming to hear us today. I’m Anne Seymour, and this is my colleague John Mark Ockerbloom, and we are the co-leads on our Penn VIVO and Symplectic Elements implementation. If you attended our colleagues talk earlier today you heard about how we got into VIVO and how it fits into our vision of a research "ecosystem" so we won’t go into details about that. I will just say that we released VIVO to our school of medicine in Jan. 2013 and went live in July. And we are early in our implementation of Elements.
How did we get 200,000 publications?

- About 4000 faculty members of different types in Penn’s Perelman School of Medicine (PSOM)
  - Research, teaching, clinical, part-time, retired
- Publications throughout career recorded in local system (“FEDS”) for promotion, tenure, evaluation
  - Everything from peer-reviewed papers to talks to op-eds
- More than just citations
  - FEDS data supplemented with abstracts & other info from PubMed

How did we get 200,000 publications? We have about 4000 faculty members of different types in Penn’s Perelman School of Medicine. Some are research faculty, some are teaching, or clinical, or part-time, or emeritus, or retired. I can safely say that 100% of these faculty have published at least 1 article and some of the most productive faculty have published well over 500 items.

Publications throughout the career of our faculty members are recorded in local system called “FEDS”. FEDS was initially developed in 2000 as a research profiling system, but without applications like VIVO it became more of a tool for promotion and tenure, through CV production. Faculty have mandates to update their records when they’re up for promotion or appointment. FEDS is also used for population of websites.

FEDS has everything from peer-reviewed papers to talks to op-eds. In our VIVO instance, we take FEDS data and supplement it with abstracts and other information from PubMed.

Elements will drive this number over 200,000 as we build comprehensive sets of publications for all faculty, not just those up for promotion and appointments.
Why read 200,000 publications? We read publications in the VIVO context to answer questions related to these publications. Questions vary by audience:
- Local scholars and their colleagues
- Administrators and tenure & promotion committees
- Librarians
- General knowledge seekers
- Funders and policy-makers

More than one mode of reading these publications:
- Some questions may require moving between different modes
- Some of these modes were inspired by digital humanists on our campus

Why read 200,000 publications? We read publications in the VIVO context to answer questions related to these publications, and many of you in this audience have a set of questions that you want to answer or have answered, often through very sophisticated data visualization.

And you know that questions will vary by audience: Local scholars might ask “Who has expertise in an area outside my discipline?” A director of our center for evidence based practice could ask: “Who at Penn can help update a Clinical Practice Guideline in the area of pancreatic cancer?”

Administrators and tenure and promotion committees might have questions like “Who at Penn is publishing high impact articles in Genomics?” Librarians might want to know “What journals are faculty in the School of Dental Medicine publishing in?”

Funders, policy-makers, and general knowledge seekers have questions too.

[John takes over the talk around this point.] There’s more than one way of reading a lot of publications. On our campus, digital humanists have helped us understand different modes of reading publications, and how people move between these modes for different purposes. Two important modes of reading are known as “distant reading” and “close reading”.

Why read 200,000 publications? We read publications in the VIVO context to answer questions related to these publications. Questions vary by audience:
- Local scholars and their colleagues
- Administrators and tenure & promotion committees
- Librarians
- General knowledge seekers
- Funders and policy-makers

More than one mode of reading these publications:
- Some questions may require moving between different modes
- Some of these modes were inspired by digital humanists on our campus

Why read 200,000 publications? We read publications in the VIVO context to answer questions related to these publications, and many of you in this audience have a set of questions that you want to answer or have answered, often through very sophisticated data visualization.

And you know that questions will vary by audience: Local scholars might ask “Who has expertise in an area outside my discipline?” A director of our center for evidence based practice could ask: “Who at Penn can help update a Clinical Practice Guideline in the area of pancreatic cancer?”

Administrators and tenure and promotion committees might have questions like “Who at Penn is publishing high impact articles in Genomics?” Librarians might want to know “What journals are faculty in the School of Dental Medicine publishing in?”

Funders, policy-makers, and general knowledge seekers have questions too.

[John takes over the talk around this point.] There’s more than one way of reading a lot of publications. On our campus, digital humanists have helped us understand different modes of reading publications, and how people move between these modes for different purposes. Two important modes of reading are known as “distant reading” and “close reading”.

Why read 200,000 publications? We read publications in the VIVO context to answer questions related to these publications. Questions vary by audience:
- Local scholars and their colleagues
- Administrators and tenure & promotion committees
- Librarians
- General knowledge seekers
- Funders and policy-makers

More than one mode of reading these publications:
- Some questions may require moving between different modes
- Some of these modes were inspired by digital humanists on our campus

Why read 200,000 publications? We read publications in the VIVO context to answer questions related to these publications, and many of you in this audience have a set of questions that you want to answer or have answered, often through very sophisticated data visualization.

And you know that questions will vary by audience: Local scholars might ask “Who has expertise in an area outside my discipline?” A director of our center for evidence based practice could ask: “Who at Penn can help update a Clinical Practice Guideline in the area of pancreatic cancer?”

Administrators and tenure and promotion committees might have questions like “Who at Penn is publishing high impact articles in Genomics?” Librarians might want to know “What journals are faculty in the School of Dental Medicine publishing in?”

Funders, policy-makers, and general knowledge seekers have questions too.

[John takes over the talk around this point.] There’s more than one way of reading a lot of publications. On our campus, digital humanists have helped us understand different modes of reading publications, and how people move between these modes for different purposes. Two important modes of reading are known as “distant reading” and “close reading”.
The term “distant reading” is the title of a book by Franco Moretti, director of the Stanford Literary Lab. What his lab does is analyze large corpuses of texts, like the millions of books that Google and other projects have digitized, to see what patterns can be discerned from computer-assisted analysis of the corpuses as a whole.
“Bird’s eye” views

It’s like the bird’s eye view of a landscape. From the point of view in this picture, you can’t see lots of detail about the buildings and grounds below, but you do get lots of information about the kind of landscape you’re flying over, its overall characteristics and layout, and features that might be worth a closer look.

“Distant reading” is meant to contrast with the more traditional scholarly concept of “close reading”, where a particular text or claim is placed under an analytic microscope to see how it works, or if it works at all.
Close reading is more like the point of view of the cat in this picture, staring intently at one bird on the ground.

Just as scientists have found both the microscope and the telescope important tools for understanding the natural world, both distant and close reading are important tools for answering questions about the work of our scholars—especially when they’ve produced hundreds of thousands of publications. VIVO can do a lot to support both distant and close readings of this body of work. But you have to think carefully about how it can best do this in your institution.
At our institution, and I imagine at many others, a lot of the people we show VIVO to are very interested in understanding the effectiveness of faculty research, and in ways that this effectiveness can be improved. Penn Medicine’s strategic plan, which includes this little graphic, stresses the importance of the “innovation, integration, and impact” of its research. This interest can be expressed in two related but distinct questions: First: What is the quality of a body of research? (Does it answer important questions in areas of interest, is it sound, is it convincingly reproducible, etc.). Second: What is its influence? (Has it changed the way people think, or act, concerning its subject? Has it connected ideas and practices in new and useful ways? Has it made people smarter, or healthier, or more efficient? Etc.)

These are both important and complex questions. For the most part, they can’t be answered directly by some single quantitative measure. But we do need to provide answers nonetheless.

The temptation, unfortunately, is often to rely on short-cuts. A lot of us in higher education have a love-hate relationship with the US News university rankings, which purport to tell you which schools and departments are superior or inferior to which others. We often say that the ranking numbers don’t really mean anything important, and also that they distort priorities in higher education. I’m not sure that both of these statements can be true at the same time.
Similarly, I’ve often heard how important it is for faculty to publish in journals with high Impact Factors. An Impact Factor is a single quantitative measure of a journal as determined by a particular commercial firm, and it’s often treated in the same way as US News rankings: even while making disclaimers about it, people often use it in practice to judge the worth of the research published in journals. This is a pernicious enough trend that over 9,000 researchers and organizations signed onto the recent DORA Declaration warning about the distortions and abuses of Impact Factors, and urging people to use other measures of the “value and influence of specific research outputs”.

Impact Factors Considered Harmful

“Do not use journal-based metrics, such as Journal Impact Factors, as a surrogate measure of the quality of individual research articles, to assess an individual scientist's contributions, or in hiring, promotion, or funding decisions.”

“Challenge research assessment practices that rely inappropriately on Journal Impact Factors and promote and teach best practice that focuses on the value and influence of specific research outputs.”

-- DORA Declaration, May 2013
http://am.ascb.org/dora/
At Penn, we’ve found that any single quantitative comparison of faculty members in systems like VIVO can be politically fraught, especially if it looks like it could be used in promotion, tenure, or salary decisions. If our VIVO shows a temporal graph of publications or grants, like the one here, should it be counting all of the ones in our system? Or just some kinds? Or should different publications and grants be weighted differently in these graphs? This is one of the reasons we don’t currently show comparison graphs in Penn VIVO. If we do show them in the future, we’ll want to make sure we support a wide range of comparisons that represent the interests of a variety of audiences.

Likewise, if you want to use VIVO to assess publication quality and influence, you’re going to want to provide a variety of ways to assess that.
So, yes, you may need to characterize the venues in which research is published because faculty care about that. But you may want to think about what data and metrics make the most sense for assessing these venues. It might be some kind of impact factor equivalent, but it might also be data about whether the venue is peer reviewed, what topic areas it covers – that’s essentially what VIVO’s Map of Science visualization relies on—or it might even be local judgments of venue quality made by your own research community. VIVO doesn’t currently support all of this data natively, of course, but if it’s important to your community, you can extend its ontology and create appropriate tools to record and analyze this data.

You can also analyze how many of the publications have associated data or implementation code provided. In some fields, providing comprehensive data is quickly becoming as important for establishing credibility as providing a comprehensive list of references. As many of you know, ANDS has created some useful extensions to VIVO to describe data sets alongside publications.

And linking publication information with assessments of that publication by other experts can also be helpful in understanding a publication’s merits. I don’t know of anyone currently doing this sort of linking in VIVO, but you see things like it in databases of book reviews and legal precedents.

<table>
<thead>
<tr>
<th>Assessing research quality</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Where did work appear, and what does that imply?</strong></td>
</tr>
<tr>
<td>· (How to best express or measure quality here? How relevant is it?)</td>
</tr>
<tr>
<td><strong>Does it have supporting data &amp; code that can be used to verify it?</strong></td>
</tr>
<tr>
<td>· How can I find it?</td>
</tr>
<tr>
<td><strong>Are public assessments of the work available?</strong></td>
</tr>
<tr>
<td>· From whom, and where?</td>
</tr>
<tr>
<td><strong>Can I review it in detail myself?</strong></td>
</tr>
<tr>
<td>· Do I have the time? Do I have access?</td>
</tr>
</tbody>
</table>
Assessing research influence

- What citations are publications getting?
  - Counts can be useful, details more useful
  - Different kinds of papers have different citation patterns
- What effect have they had on practices?
  - Patents, procedures, products, clinical outcomes, education
- What impact is shown via alt-metrics?
  - (and which are irrelevant, or need to be monitored for gaming?)

- Are publications and their supporting material openly and widely available?

You can also use data to see how publications are being cited. Some VIVO instances are including citation counts with their publication data. If you want to compare counts for different publications, though, it’s helpful to track what kinds of publications they are, since citation patterns can be quite different for surveys than for original research findings, or for papers that try to replicate or critically evaluate previous findings. Penn Medicine’s internal publication database already categorizes some of these different kinds of papers, and we’re considering ways to better represent these categories in our VIVO service.

The influence of a publication can be also be assessed by how it changes practices. Linking a publication to related patents, clinical procedures, products, or educational materials can be a powerful way of showing the impact of the research. VIVO can already model a number of these concepts, and can be extended to model more. You may want to consider whether and how you want to link them to publications in your VIVO instance. Various kinds of “alt-metrics” can also help gauge influence, but you’ll want to consider carefully which ones are most useful.

Another question you might want to ask is “Are publications and supporting data openly accessible?” Because that in itself increases their impact potential.
Beth Noveck’s keynote yesterday mentioned the challenges of determining the safety and effectiveness of medical devices. Well, here’s a recent headline of one story showing how open access to data helped researchers find out whether a medical device was more effective than existing alternatives. (Spoiler: No.)
There have been other well-reported cases of the impact open data has made in fields ranging from genetics...
...to economics.

But the important take-away, again, is that if you want to support research evaluation, you have to support many different ways of assessing the research.
Not only do you need a variety of measures to assess the publications of any particular kind of researcher, but you also need to support different kinds of assessments for different kinds of researchers.

At Penn, for instance, our clinical medical faculty are assessed not just on papers they author, but on papers where they conducted supporting studies and trials, even when they don’t appear on the published author list. The publications that do bear their names are often conference presentations and various kinds of gray literature that often don’t show up in medical publication databases. We need to make sure these publications, and these contributor relationships, are properly accounted for in our VIVO data. Similarly, other disciplines like computer science, anthropology, or the arts may have different ways of evaluating, or even representing, their publications.

And of course if you really want a deep understanding of a scholar’s publications, there’s no substitute for taking a close look at some of it. VIVO can help you with that sort of close reading too.
In particular, VIVO can help with two important requirements for close reading: Finding the things you need to pay close attention to, and then getting as good a view as you can of those things.

Effective search is an essential element of finding relevant publications, and the more you can put into a publication’s record that relates to what it’s about, the more effective searches in VIVO will be. That’s one reason we aim to add abstracts and other subject-related terms to our publication records in VIVO wherever we can. Currently we’re using Pubmed to do this, and we hope to have Symplectic’s Elements help us with that going forward.
Our faculty also find it important to designate some of their publications as especially worthy of attention. This can also be a useful way of guiding close reading. Out of the box, VIVO doesn’t have a way that I know of to highlight selected publications in a larger set, but there are various ways you can extend VIVO to do this. Hahn Ryan at UC San Francisco, for instance, has developed an OpenSocial gadget for highlighting selected publications.

http://open-proposals.ucsf.edu/opensocial-gadgets/proposal/503
VIVO can also support close reading of publications by making it as easy as possible to get to the content and supporting materials of those publications. Links to full text, including institutional repository and other open access versions, can be very useful here. Duke’s added such links to its VIVO service, for instance, and we’d like to add them too. It’d be great if these links were an explicit part of the VIVO ontology.

Looking closely at the supporting materials for a publication, or other work that responds to the publication, can also be useful, as I’ve already mentioned. Analyzing the response to a widely-cited publication may involve shifting from close reading back to distant reading, of course.
There are many other useful distant reading related questions that VIVO can answer, and it often answer them directly if the right data is included. For instance, you might want to know whether your researchers are complying with funder requirements for open access, or data provision. Drew Wright and Paul Albert at Weill Cornell have done some good work here on using VIVO data to assess NIH open access compliance, and their poster is included in this year’s program.

People on hiring committees might want to know what subject areas current faculty are publishing in, any notable strengths or gaps in subject areas, as well as recent research trends, to help determine where new faculty members might strengthen a program.

We in libraries are quite interested in knowing what venues our faculty are publishing in, and how that has changed over time, so that we can make better collection decisions to support their research.

And, of course, there are all the author network questions that VIVO is especially well-known for. I won’t reiterate them all here, but they represent an important kind of distant reading of our publications.
Conclusions

- **VIVO should help answer questions about scholars’ research through distant and close reading of their publications (and other data in VIVO)**

  - Getting right data into VIVO important
    - Consult your constituents to find out what “right data” is
    - Extend VIVO data, queries, and/or interfaces appropriately

  - VIVO has useful features for modeling this data
    - E.g. link to grants and patents, researcher networks…

  - VIVO could use some more of them
    - E.g. full text links, selected “highlights” for review…

So, in this short time we’ve had, I hope we’ve given you some ideas about some of the useful questions you can answer about scholarly research through distant and close readings in large set of publications. We’ve shown you how a VIVO service can support useful distant and close readings, if you give it the right data and make it easy to use the data to answer your questions. You’ll need to work with your intended audiences to determine the questions and the data that are most important to them. And you’ll probably find ways to support some of these inquiries in VIVO as it exists today, but you might also need to extend VIVO further to better support the kinds of readings your audiences want.
Questions for you

- What sorts of questions about publications (and other research activity) do you want to answer with VIVO?
- What “distant” or “close” reading mechanisms are (or would be) most useful in your VIVO installation?
- How do your various audiences influence the kinds of reading you give priority to supporting?

- Do you want to talk with us more?
  - John Mark Ockerbloom: ockerblo@upenn.edu
  - Anne Seymour: seymoura@upenn.edu

So, with that in mind, we’d like to leave you with some questions to consider about what kinds of inquiries and readings you find most important for your VIVO service. Or in interesting ways you could support them. And if any of you want to talk to us afterwards, our email addresses are on this slide. Thanks!