Apps and Tools Call 20141021

Date	2014-10-21
Topic	Mountain West Research Consortium

Hackathon recap

- http://www.vivoweb.org/blog/2014/10/fall-2014-vivo-hackathon-report
- https://drive.google.com/folderview?id=0B6osQY8NM9BGNHA0cmpiaUE2VzA&usp=sharing
- Thinking of a 1-2 day virtual hackathon as a follow-up, using a Google Hangout or GoToMeeting or WebEx
    - one possible date(s): Monday &/or Tuesday before Thanksgiving (Nov 24-25)

Planning Spring, 2015 VIVO Implementation Fest

- Portland, OR -- hosted by Oregon Health & Science University
- We'd like to include as many west coast schools as we can
- potential dates: March 16 to 20th -- are there any conflicting major meetings?

Eric Garbin from Univ of New Mexico

- Mountain West Cross-Consortium Search (https://search.vivo.ctr-in.org/)
    - a group of affiliates in clinical and translational science (a CTSA)
    - goes beyond VIVO
    - listed in the direct2experts search site as well
    - host 11 VIVOs and ingest the data for and host 10 of them
        - many of them are still experimenting with how they will run VIVO and who will be locally responsible
    - University of New Mexico is the largest, with University of Idaho and University of Hawaii next
        - Idaho is run by the main library so not as specifically geared to health sciences -- and they are running 1.6
        - Hawaii is in part an ancestral Harvard Profiles beta instance (not RDF at all, but XML, so had to write custom XML transforms to get that data into Solr) and part VIVO
- Search built on the 2011 http://vivosearch.org work using Scala actors framework for concurrent threads in the crawler, once URLs are discovered -- the "linked data indexer code"
    - https://github.com/vivo-project/Linked-Data-Indexer
    - There is also 2013 work that uses more of a Hadoop approach
    - There differences in how different types are used in different VIVOs that they hope can be made more consistent
        - Also differences in how Idaho represents research areas from what New Mexico does
    - The indexer does a complete re-index since it doesn't have the capability of deleting records from the Solr index
    - Has issues:
        - scaling this type of harvesting to more kinds of data within the VIVOs, going beyond faculty and research areas
        - addressing search result rankings in the aggregate vs. the rankings when search each individual site
- Questions?
    - Layne: do you meet with colleagues from the other sites? Yes, have a manager/supervisor at UNM in the informatics department who has been identifying points of contact to help set up faculty profiles for the different universities
    - Jim: worked on the 2013 effort you mentioned so I have some idea of the previous version -- congratulations. Our beta site was harvested once completely and it's great to see something in production.
        - Eric -- was a learning curve to the Scala actors framework and challenges setting up the maximum number of connections
        - Another developer set up the Drupal theme -- modifying the theme that Miles Worthington had originally set up and updating it to work with Solr 4
        - Other tweaks to update in Windows
        - Added a research area field in Solr and boosted that since faculty do tend to add research areas
        - We wanted to reproduce a previous effort and ramp up on the concurrency issues with indexing more gradually -- if you can fix something small, you can learn a lot in the process from the framework without having to understand everything before you start
        - Solr is a good choice because that enabled us to index the University of Hawaii Profiles data
    - Tenille Johnson (eagle-i) -- the Profiles instance at Hawaii has pulled in information on cores at any of the RCMI sites and research attached to them -- do you have interest in expanding the data in your search to include resources?
        - The site at http://www2.rtrn.net/researchhub includes an integration of Profiles with data from eagle-i -- a tool they built themselves to re-purpose eagle-i data (in RDF). Also data for University of Alaska Fairbanks and the University of Montana
        - Eric -- ideally we'd be populating the VIVO instance from Profiles to save them the work to re-enter it
        - The Profiles site at Harvard returns RDF; there is also a new version of Profiles (2.5.1), running at Harvard, that includes a built-in hook for eagle-i -- and will pull in data from eagle-i into the person's Profiles page
    - Jon -- indexing publications?

        - Eric -- don't have the data yet
        - There is an opportunity to relevance ranking when the data are harvested, but that has to be worked out given the different approaches to research areas
    - Alex -- shared a link to a profile in Harvard Profiles that pulls in eagle-i data
        - Direct2Experts?
        - Eric -- happy to participate in it but wasn't really enough
    - Jon -- do you think you can you come to agreement on the level of detail in types, or
        - most of the research areas are just owl:Things
        - believes it is handled better in VIVO 1.6
    - Jon -- problems distinguishing common names from different institutions, where each may have a Department of Medicine or Department of Pediatrics
    - Eric -- the University of New Mexico has it's own instance, as does Idaho, but the other 10 are in a single Tomcat
        - not yet reconciling a person who is in two different VIVOs
        - are asserting co-authors not at the primary institution only as foaf:Persons, and they do not get pulled into the shared search
- Building on Vivo Harvester for data entry and cleaning
    - Building a publication graph pulling from PubMed
        - specify the names in VIVO in a PubMed-friendly way, using only 1 token for the first name or initial
        - developed Python scripts to handle the RDF editing to clean up the data -- e.g., a deduplication script and a name linking script that incorporates rules that work better than the default Harvester rules, which are very conservative so result in a lot of extra entities being minted
            - have a GUI for the tool but haven't yet plugged in the scripts
        - also identify PubMed collaborators in their VIVO at UNM -- requires further disambiguation, including dealing with two entries for the same person as author on the same publication
        - working to de-couple the different processes
    - Also have an online Vivo Ongoing Refinement (VOR) Tool
        - can trigger new harvests depending on different departments so can update on a more regular schedule
        - in the process of re-working it to work for other hosted VIVOs
        - Chris -- is the code up on GitHub yet? Not yet but has started setting it up
        - Eric -- works quite well with the Harvester process, where the interim results are output files
            - Having talked with the University of Idaho where they use Google Refine for some of the harvesting and translating, and that fits into the overall Harvester process
            - Wants to get that all working in 1.5 before moves to 1.6
            - Chris -- some change a lot -- the FOAF and vCard thing is mostly where it changes
            - Eric -- the vCard structure might be an advantage since you don't have to create a foaf:Person for non-faculty members that you don't otherwise know anything about
                - can take from email attachments -- are thinking of email surveys to faculty members from VIVO, where can pre-populate the email with the person's title and publications, and have them answer to confirm
                - had a prototype that ran last year for the Department of Medicine
                - that could cut down on a lot of RDF editing they might have to do
            - Chris -- would like to collaborate with you on sending emails with suggested pubs, where people can click a hyperlink to indicate yes or no
                - Eric -- cuts down on the amount of name disambiguation you have to do after the fact
                - If you set up the department by URI it will set up harvest jobs for all the faculty in that department

- Thanks to Eric for presenting!

November 3rd ontology call -- in 2 weeks