Implementation and Development Call 20131003

Calls are held every Thursday at 1 pm eastern time (GMT–4 in daylight savings, GMT-5 standard time) – convert to your time at http://www.thetimezoneconverter.com

View and edit this page permanently at https://wiki.duraspace.org/x/JfcQAg, or use the temporal Google Doc for collaborative note taking during the call.

VIVO is hiring!

DuraSpace is seeking a dynamic and entrepreneurial Project Director for the open source VIVO project (www.vivoweb.org), a world-wide community focused on creating software tools, ontologies, and services. The VIVO Project Director will have the opportunity to play a major role in a collaborative movement that will shape the future of research.

See full posting – applications are scheduled **to close on or near October 23rd**. Note that there is no requirement to be a U.S. citizen.

Release update

No release candidate has been created yet – progress each day.

Apps and Tools Group

Notes from Sept. 24 meeting recorded as this webcast showing a Python data checker for VIVO developed at the University of Florida.

Next meeting on Tuesday (October 8) at 1pm Eastern – Stephen Williams from the University of Colorado will host. Stephen has posted an agenda to the vivo-dev-all mailing list.

Paul -- great that you are recording the sessions and posting them to YouTube.

Upcoming Events

- 2nd Annual CASRAI International Conference, October 16-18 in Ottawa
  - Conference streams: Reconnect Big Data, Reconnect the Library, and Reconnect the Machine
  - http://reconnect.casrai.org
  - **Jon will be presenting on VIVO, along with Memorial University**

- 1st Annual UCosmic Conference, October 31 in New York
  - **Collaborative Software Development to Address Strategic Challenges in Higher Education: Kuali, VIVO and UCosmic**
  - http://www.ucosmic.org/Conference.aspx

Updates

- Brown (Ted)
  - finalizing import of data from existing research profiles
  - hope to have a public release date set soon
  - created a Vagrant bootstrap script for VIVO. Will install VIVO on an Ubuntu server image.
    - https://github.com/lawlesst/vivo-vagrant
- Colorado (Alex & Stephen) In the middle of Elements curation and behind on listserv responses
  - Working on Elements publications curation for 2013
  - Stephen will be catching up on VIVO emails after recovery from flooding
- Cornell (Jon, Jim, et al.)
  - 1.6, 1.6, 1.6, 1.6, 1.6…
- Duke (Richard)
  - reloading grants data from our source system. we put grants into their own graph and then wipe/reload that during a full grant load process. most days we just do an incremental load.
  - search re-indexing process taking a really long time >5h, sometimes ~1.5h -- doesn't seem to correspond to the number of new triples -- looking forward to incremental re-index
    - UF had an issue with bad characters taking a long time to fail
    - new version of Solr in latest VIVO repo
    - Jon -- any correlation to inferencing? Richard indicated no, not running inferencer as they ingest all triples to not require it -- using a Ruby script that they could possibly extract and share
- Florida (Chris)
  - Had a second successful run of people ingest from people soft
    - Developing a weekly process
    - Deploying ingest from git repository
  - Working on visualizations with d3js (http://d3js.org/) and JSON -- which can be generated from within VIVO -- Javascript visualizations seem fast! Probably demo in last October Tools call.
- NYU (Yin)
  - talking to production group about graduation project -- been working as a dev/research project
  - been using an intermediate data format for getting data into VIVO, but prod group wants to connect VIVO (?) to enterprise data warehouse -- are there best practices for this? Jon clarified if they want a realtime connect vs ETL -- suggested that the closer the transformation gets to RDF, the easier it is to bring it into VIVO -- Ted happy with Python and RDFlib, UF been using Python and starting to use RDFlib

    - https://github.com/ufvivotech/ufDataQualityImprovement/tree/master/vivotools
    - https://github.com/nrejack/dchecker
    - question about not using front end, rather back end RDF via XML or URLs, and Solr search?
    - Also suggested VIVO hardware requirements? Chris suggest AWS specs on wiki.
        - AWS Specs for UFL VIVO Hosts:
        - X-Large-Memory
            - 17.1GB (ram)
            - 6.5 EC2 (cores)
            - 420GB
            - 64-bit moderate
            - m2.xlarge
        - 2X-Large-Memory
            - 34.2GB (ram)
            - 13 EC2 (cores)
            - 850GB
            - 64-bit high
            - m2.2xlarge
- Scripps (Michaeleen)
    - Stella has a working version of the grants ingest from NIH Reporter. Ingest program written for 1.5.1. Not sure if she should share for that reason?... Jon: it would be helpful to post regardless!
    - Stella is also working on authorship representation.
    - Representing patents
- Stony Brook (Tammy)
    - Using JSON to integrate at data interface between Java and Python dev efforts
- UCSF (Eric)
    - Bringing in grants from NIH Exporter -- Jon mentioned concern of annual updates to long running (25y) NIH grants -- Stella's looked into how to best represent these in VIVO ontology
    - Author registry idea; would be compatible with ORCID and include ORCID ID -- aim for lower policy hurdles
    - Anyone look at Project Honeypot tools to keep bad traffic away from site? HTTP Blacklist catch around 10k HTTP requests per day. UF also blocked access to CPU heavy pages like the visualizations, for web spiders that don't honor robots.txt.
- Weill Cornell (Paul)
    - reconciling self-reported publication data with data from VIVO instance -- very few pubs rejected, many were duplicates already in VIVO -- Ted offered some good advice
    - template updates

Notable list traffic

See the vivo-dev-all archive and vivo-imp-issues archive for complete email threads

1. PubMed Harvester doesn't like particular records (Lynda, Andy)

- The Harvester is essentially unusable with PubMedFetch, due to bugs in code from NIH. Some records in PubMed have data which is not correctly handled by the NIH code. It's possible to work around these bugs by using PubMedHTTPFetch instead of PubMedFetch. However, you need to URL-encode your search request if using the HTTP version.

2. ExternalAuthId and named graphs – fixed by Jim and tested by Ted VIVO-305 - Matching property should work if the triple is in a named graph (not just kb-2) - RESOLVED

3. VIVO and TDB (Michel, Ted, JohnF) – have VIVO working with a TDB database, instead of SDB (so a relational database back end store is no longer needed)

- Ted: Fuseki 1.0 was released last week and I was able to get that to connect to an instance of VIVO 1.5 using the same endpoints you specified:
    - VitroConnection.DataSource.endpointURI = http://localhost:3030/tdb/sparql
    - VitroConnection.DataSource.updateEndpointURI =http://localhost:3030/tdb/update
- Michel: I now want to write a java program with Jena, where I insert data into the TDB. I want to use the Jena api, with model.createResource and resource.addProperty and so on.
- Ted: I use Python and RDFLib [1] for VIVO data loading. RDFLib, as of version 4, supports SPARQL 1.1 so you could use that to write directly to Fuseki
    - As for learning about the VIVO ontology, one technique that I've heard recommended and find useful is to use the VIVO admin to create resources that you want to load (FacultyMember, Book, etc) and then inspect the RDF that is generated to see how the data is modeled. VIVO will serve Turtle for a resource (e.g. n1234) by pointing your browser at http://localhost:8080/vivo/rdf/n1234/n1234.ttl.
- JohnF: Specifically, take a look at the org.vivoweb.harvester.utilrepo.JenaConnect class. It's an abstract class that is extended by SDBJenaConnect and TDBJenaConnect. It should give you a good idea what you'll need to do to insert RDF into VIVO using the Jena API.

4. Finding grants via investigator name (Michaeleen) – having an investigator relationship with a grant is not sufficient to make the grant show up in the results for a search on person name.

5. Uploading image when editing a property of an individual (Yu, Huda, JohnE, PatrickW, BrianC)

Call-in Information

Date: Every Thursday, no end date

Time: 1:00 pm, Eastern Daylight Time (New York, GMT-04:00)

Meeting Number: 641 825 891

To join the online meeting

Go to https://cornell.webex.com/cornell/e.php?AT=WMI&EventID=167096322&RT=MiM2

If requested, enter your name and email address.

Click "Join".

To view in other time zones or languages, please click the link: https://cornell.webex.com/cornell/globalcallin.php?serviceType=MC&ED=167096322&tollFree=1

If those links don't work, please visit the Cornell meeting page and look for a VIVO meeting.

To join the audio conference only

To receive a call back, provide your phone number when you join the meeting, or call the number below and enter the access code.

Call-in toll-free number (US/Canada): 1-855-244-8681

Call-in toll number (US/Canada): 1-650-479-3207

Global call-in numbers: https://cornell.webex.com/cornelluniversity/globalcallin.php?serviceType=MC&ED=161711167&tollFree=1

Toll-free dialing restrictions: http://www.webex.com/pdf/tollfree_restrictions.pdf

Access code:645 873 290