2011-10-06 Development Leads Call

Updates

- Cornell
  - migrating from one physical server to 2 virtual ones, separating the database from the app – this enables us to use the standard virtual server provisioned
- Duke
  - preparing to pull in data from an existing system, including HR and publications. Looking to do courses after that
- Florida
  - photo upload through dev and staging and looking at adding 27K images to production very soon
  - moving to EC2 instance – calculated the data charges. Conceptually very interesting to have separate SPARQL server from VIVO website, and will have RDS databases that can spin up and synchronize clones of the database. Throw up development and staging servers on an ad hoc basis.
    - have to consider how to get data from institutional sources to the cloud – could be either policy or performance issues
  - Joseki config has been updated; running under Jetty and want to move to Tomcat – run currently on the same server as VIVO but the database is on a different server. Doing replication of the database itself so a pathological query to the SPARQL endpoint won't slow down the VIVO production service
    - John Fereira has done that and requires a minor modification so don't expose database passwords
    - wants to run on a different host
    - looking at a database connection pooling manager called c3pio
      - UF had trouble with connections timing out
- Indiana
  - Have updated the SciVal journal listings to cover 2001-2010 and hope to get that in v1.4
  - Ying wants to access the SPARQL query builder without having a
- North Texas
  - planning meeting Friday with the faculty profiling folks to lay out plans for their VIVO implementation. Will feature one research cluster.
- Stony Brook
  - Janos has been working with the CrossRef DOI service as a way to pull back publications – might want to contact Mummi Thorisson since he's also worked with it. We can check with Geoffrey Bilder about permissions to harvest data.
- Weill
  - upgrading to 1.3 in next couple of weeks
  - working on disambiguation of Scopus data
    - looked at Griffin Weber's tools but they seem to require Microsoft SQL Server
- Wisconsin

Special Topics

- John Fereira has volunteered to go over the work he's done with the JSON and XPATH harvesters he's been working on, and can demo the JSON harvester. This works builds on the UF Harvester code.
  Working with the international ag community for some years; will be in Rome next week at the FAO. Not on the VIVO team but have done a lot of integration work with VIVO. Recently at a conference in Beijing working on one of the agriculture systems at FAO. One is called AgriDrupal, an agriculture information systems customization of Drupal that is available for download. Have done the same thing for DSpace, and John has proposed an agriVIVO with ag-focused capabilities built in.
  Next week looking with FAO at pulling data from existing systems for ag researchers, run by different organizations, with some overlap but many people represented only on one.
  If you have a Drupal intallation, you can install a module that will output your content types as JSON, and have looked at harvesting from this Drupal VIew to VIVO.
  CIARD Ring with 150 organizations, represented in a structured JSON view. Wanted something that could fetch this content and go through a series of harvesting steps.
  Also wanted to run on Windows so set up copies of the shell scripts as PHP scripts so could use a web interface. Mimics the BASH shell, and invokes a PHP version of the scripts from the bin directory of the Harvester – more cross-platform mechanism. Janos – Cyqwin might be able to run BASH scripts. Really just a bunch of system calls.
  Created a JSONFetch class based on the JDBCFetch class, but am connecting to a URL or from a file instead of a database connection.
  Pull in the JSON as a string; working with a name, id, and path for each node used for pulling person nodes out of the JSON (e.g., $..users.. users). Uses the description field to differentiate organization nodes from person nodes in the data when move on to RDF generation.
  From that point on it follows the same path as the rest of the harvesters – transformations with XSLT, goes on through the scoring and transfer steps that get you to a VIVO instance.
  Hoping next week to harvest 6500 people from one site and 800 from another next week.

Since using the JSONPath realized it would be pretty similar to harvest using XPath – for fairly complex XML in multiple namespaces, in a common agriculture metadata format.

One of the limitations now with the JSONFetcher is that it's not recursive – requires a flat JSON model, and would like to go back to improve it after this trip to Rome. The XPath fetcher does have the ability to recurse into the data structure.

Got a nice demo today from Huda Khan about the use of the terminology annotations from the Stony Brook UMLS service as demonstrated at the VIVO Conference. Might be able to invoke during a harvest to resolve terms in incoming data to existing controlled vocabulary URIs.

John's semantic services could store the SPARQL query (e.g., to DbPedia), convert to JSON, and the fetch.

Is a JSON-LOD project that represents LOD as JSON.

Notable Development List Traffic

- Question from Cliff: Are there any services available in VIVO that allow soap/rest interaction?
  - Nick Skaggs' response: *We have some examples of PHP interacting with VIVO data on the downloads section of the sourceforge site — one that interacts with a wordpress site and a drupal site. They are simple examples. The idea behind linked open data is that the data in consumable and self-describing. That said, you could attach something in front of the RDF store and serve up the rdf not as RDF but in a JSON format or something similar. We (The University of Florida) serve our VIVO data publicly thru a sparql endpoint in addition to the website at sparql.vivo.ufl.edu. Via this interface you can use a library like ARC for php (which the examples on the site use I believe), but there are others.*
  - John Fereira's response: _I would also recommended looking at ARC2. I don't know if you're using Drupal for your interaction but even if you're not you might want to look at the Feeds rdf_importer module that Miles Worthington developed (http://drupal.org/sandbox/milesw/1085078) as it uses ARC2 for getting data into PHP and has been tested quite a bit with VIVO. I have also developed a java web app that issues sparql queries against an endpoint (works with either a Sesame or Joseki endpoint). The results are transformed to a native Java object that can be serialized in multiple formats include JSON or XML or the original rdf/xml. I've used the JSON output to create content (from VIVO) in a Drupal (PHP) instance and the XML output to created content in a Cold Fusion based CMS. The web app has a small database for storing "canned" sparql queries. The queries can be "parameterized" so that, for example, the same query can be used return a list of faculty from the Chemistry or Linguistics department by passing in the appropriate department id. The service is available at: http://sourceforge.net/projects/semanticservice/develop_
- VIVO on Ubuntu
- date formats in CSV files
- VM documentation
- SPARQL query generator

Call-in Information

1. Please join my meeting, Thursday, October 6 at 1 pm EDT https://www1.gotomeeting.com/join/431156241

2. Use your microphone and speakers (VoIP) - a headset is recommended. Or, call in using your telephone.

Dial +1 (630) 869-1011
Access Code: 431-156-241
Audio PIN: Shown after joining the meeting

Meeting ID: 431-156-241

last meeting | next meeting