2012-01-05 Development Call

Please add additional agenda items or updates

Updates

1. Colorado
2. Cornell – *upgrading to 1.4 this week*
3. Duke *(can't attend this week)*
4. Florida
5. Indiana
6. North Texas
7. Stony Brook
8. Weill Cornell
9. ;Continue with getting author ID and DOI/pmid from Scopus, then integrate with Pubmed Harvester for ingest into VIVO.
10. ;ReCaptcha for VIVO 1.5? NIHVIVO-3538
11. Any other sites

Short Topics

- UF has fully transitioned http://vivo.ufl.edu to Amazon hosting and is using the Amazon MySQL environment that optimizes MySQL performance – and is noticeably fast.
- VIVO project pages on http://vivo.sourceforge.net have been simplified and updated lightly – feedback and suggestions for other content or reorganization are welcome
- Report on a December meeting with the Weill Cornell team on their User Stories: Defining features and functionality VIVO needs - September 2011
- Planning the 1.5 release – goals already identified, time frame, components visible in the development Jira space, and how to provide input

Notable Development List Traffic

- Ongoing discussion – Solr deployment and permissions – how much are people being tripped up by assumptions at Cornell about user accounts and group permissions?
- Wes Rood – properties exposed when logged in as the root user
- Tammy – preventing the collation of publications by subclass in order to get all pubs in reverse chronological order
- Ann Gardner – background and contributor tags
- John Fereira – harvesting events from Google Calendar and converting command line scripts to PHP

Discussion: How to improve full lifecycle data management in VIVO

See Development Call 20120105

Adding data to VIVO is not easy – partly because RDF and semantic tools are unfamiliar, but partly because of fundamental data management challenges facing any large system with multiple sources, especially when some data may be edited after ingest and those changes affect subsequent updates.

We have collectively and individually developed many different approaches for adding VIVO data, most notably the Harvester, but also extensions to Google Refine and to VIVO to support using Google Refine, and modifications to the D2R Map tool, improvements to the ingest tools in VIVO itself, and integration of some Harvester functions into VIVO. Joe McEnerney at Cornell has also worked extensively with data from Activity Insight, a faculty reporting tool used by many universities in the U.S., and in the process developed a number of practices for managing initial data matching, incremental updates, procedures for retraction of data that may have been modified in VIVO, and detection of malformed or orphaned data in VIVO.

As VIVO matures at each of our institutions, we are also being asked more questions about reusing data from VIVO in other applications, about reporting using data in VIVO, and tools for archiving, removing, or exporting what can be very large amounts of data. How can we address these challenges appropriately?

Questions from the UF Harvester team

In our discussions of ingesting Person data from People Soft we have a wish list of things we'd love to know about a triple to allow us to preform intelligent actions on any triple as part of an ingest process. Some of these are:

- CreatedBy = What user or person created this triple
- CreateDate = When was this triple first created
- LastModBy = What user or person last modified this triple
- LastModDate = When was this triple last modified
- Public = Am I allowed to show this triple to the public

Other questions to address

As time permits this week, and for future meetings

1. examples of current best practices for recurring ingest processes
2. fundamentals – strategies for when data can be replaced and updated en masse vs. updating that must allow for the possibility of end-user or curator editing
3. how to segregate data by Jena graph using SDB, and VIVO's current limitations and foibles in this regard
4. how to establish a baseline for each different data source
5. what has to be archived prior to, during, or after each successive update – how much can be done in bulk vs. statement-by-statement
6. what an incremental ingest process typically produces: RDF to add and RDF to remove
7. pitfalls of removing RDF data (it may have been introduced via other processes, and RDF does not accumulate multiples of the same triples – if you remove it, it's gone, even if two separate processes had independently added the same statements)
8. techniques for rolling back additions of RDF via SPARQL CONSTRUCT
9. what can be accomplished by logging actions and/or triples
10. what would be reasonable near-term goals for the Harvester
11. what would be reasonable near-term goals for VIVO (1.5)
12. what use case, requirements gathering, or design work needs to be done and who can participate

Next Week

Jim Blake will give a presentation on modularity and extension options for VIVO and the work he is proposing for VIVO 1.5

Call-in Information

1. Please join my meeting. https://www1.gotomeeting.com/join/322087560

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

Dial +1 (773) 897-3008
Access Code: 322-087-560
Audio PIN: Shown after joining the meeting

Meeting ID: 322-087-560

last meeting | next meeting