Data Ingest Workshop

A guided journey in processing authoritative data into VIVO

Author: Stephen V. Williams (svwilliams@gmail.com)

Goals

At the completion of this workshop, participants will have hands on experience processing tabular relational data into VIVO using the UI and then transforming that initial process into an automated process. They will be able to identify the various methods used to get data into VIVO and be ready to explore some of them on their own.

Audience

The workshops intended audience is individuals with basic to advanced knowledge of data. If you can write a simple sql query or ask for data from MS Access this course is for you.

Outline

The course will be broken into four parts with a 10 minute break

Lecture: Landscape of Institutional Data (20 min)

It's important that the participants think about the original use case for the data they have obtained before proceeding to place it into VIVO. Often this use case shapes either the workflow for getting the data in, or the curation process after it's been initially ingested. This initial lecture will highlight the importance of understanding "What does it contain", "What is it used for", "How often does it change", and "How does it reflect the perceptions at my institution".

Hands-On Tutorial: Ingesting Data into VIVO (90 min)

The participants will be provided with a sample VIVO database (in .sql) and a sample data file in CSV. The instructors, some filtering around the room to assist, will then walk the attendees through processing that CSV file into VIVO using the VIVO UI. We'll cover model basics, data import, sparql constructs, and blank node renaming, as we walk through getting our data into VIVO. We'll then take the same data and turn it into a VIVO Harvester script walking back through each step of the UI process and translating it to a corresponding Harvester process.

Demonstrations: Other Tools and Advanced Concepts (60 min)

We'll restart the workshop with the Harvester and explaining how complex an ingest can get with the Pubmed Harvest. This script uses advanced aspects of the VIVO Harvester and touches upon some of the difficulties with data sources, such as disambiguation. From there we'll move onto another tool for disambiguation, Google Refine. We'll also explore projects like Karma which intend to make getting data into VIVO simpler and some examples of institutional specific code written against VIVO. Participants can follow along, however due to time constraints these demonstrations will not wait for participants to catch up. Each demonstration will be about 10-20 minutes in length.

Group Discussion: Best Practices (30 min)

To close the workshop we intend to bring in implementers from around the VIVO Collaboration to discuss best practices in data ingest. From curation, to automation we hope that this discussion will highlight that there is no right way, but many poor ways to ingest data. Through constant refinement an institution can see stable, understandable authoritative data flowing into their VIVO system.

Materials Required for the Workshop

Each participant will be required to bring a computer (or tablet with access to a VM) with one of the following configurations

Their own environment:	**A local Virtual Machine:**	**Their own ec2 cloud**
Java, Ant, Maven, Tomcat, MySQL, VIVO, Vitro, VIVO Harvester	Oracle VirtualBox	

A month before the conference participants will be sent an email detailing the necessary materials for this workshop. The email will detail how to setup their environments for either of the three options. The email will also include a brief survey to access the participants and allow the instructors to re-tailor the workshop as necessary. 2 weeks before the conference begins a follow up email along with download instructions for the pre-built Virtual Machines will go out. Instructors will be on hand an hour before the workshop to help any participants struggling to setup their environment.