A Lot of Data for a Little Code: Get Data into VIVO Faster with the Jena Framework

Ryan Cobine - rcobine@indiana.edu
David Cliff - dgcliff@iu.edu
Robert H. McDonald - robert@indiana.edu
Jon W. Dunn - jwd@iu.edu
Robert Light - lightr@indiana.edu

INDIANA UNIVERSITY

Overview

VIVO's Advanced Data Tools in the UI and the Harvester provide two ends of the simplicity/power spectrum of data ingest. Using the bundled Jena framework with relatively simple Java terminal applications provides an option more powerful than the UI, yet lighter weight than creating a harvest from scratch.

The Challenge

Indiana University's VIVO implementation team was called upon to assist with data ingest into the SEAD project's VIVO instance. The initial task was to get as many PIs and co-PIs and their publications into the SEAD-VIVO instance as quickly as possible.

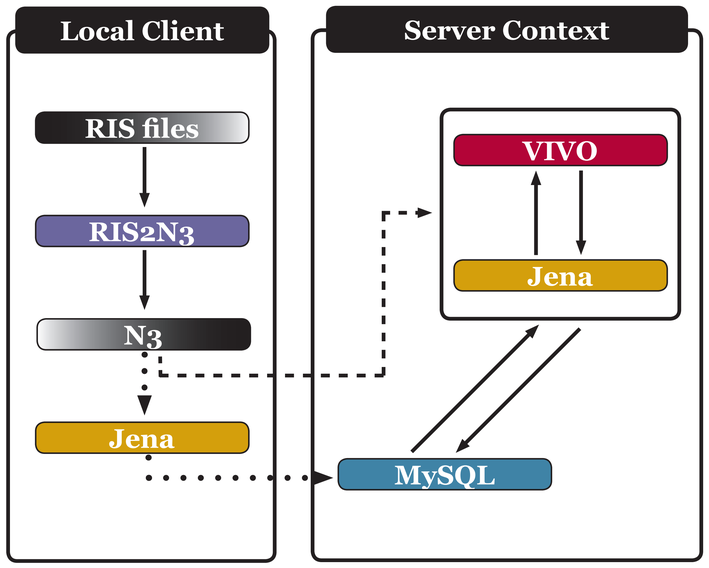
Characteristics of the Data

There was no central researcher database available, so all researcher and publication data had been assembled manually with the assistanc of a variety of utilities, and posessed the following characteristics:

- Stored in RIS file format (AKA RefMan)
- Too great in volume to reasonably enter manaully
- Too complex to ingest with a CSV-to-RDF conversion and import
- A "one off" ingest, that is, a manually assembled data source that is not maintained or programmatically updated—the overhead of creating a harvester ingest was not a good tradeoff

Figure 1: The Middle of the Ease/Power Spectrum

Figure 2: Relationship of Components



Our Solution: RIS2N3

RIS2N3 is a Java terminal application with four essential functions:

- RIS file parser
- Publication and author definition (with name disambiguation)
- Relationship definition
- N3 output (and optional loading)

Benefits

- RIS is a broadly supported and well understood file format for citation data. Google Scholar and EndNote both export to this format.
- Incorporates templates to handle different publication types (e.g. journal article, conference paper, book chapter, etc.).
- Results in the most fully qualified people names possible across the source data, while reducing the liklihood of person duplication
- Creates relationships between publications and authors (sufficient for co-authorship visualizations), and between publications and publishers
- Produces N3 file output or a direct load into VIVO

Where Did the Data Come From?

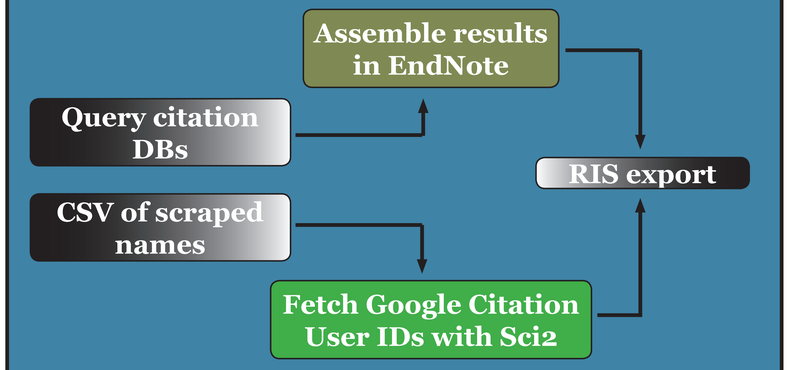
Two separate ingests were run. The first was manually assembled:

- Citation databases such as Google Scholar, Proquest, etc. were queried for each selected researcher.
- Results in various formats (BibTech, ACM, RIS, etc.) were collected in EndNote and exported to the RIS format.

The second set began with names scraped from the project website into a CSV file:

- The CSV was fed into the Sci2 Tool, which returned Google Citation User IDs for the names it matchedvv.
- The matches were manually verified, and the Citation IDs were used to identify Google Scholar profiles.
- Publication lists were exported from Google Scholar to the RIS format.

Figure 3: Two Paths to RIS Ingest



What is SEAD?

Awarded through NSF's DataNet program, the Sustainable Environment-Actionable Data (SEAD) project is developing tools and services for active curation and longterm preservation of scientific data, while also engaging researchers through social networking tools.

SEAD will enable new modalities of sustainability science—the study of dynamic interactions between nature and society by advancing the science of sustainability through the integration of social science, natural science, and environmental data.

More info about the SEAD project: http://sead-data.net/

Acknowledgements—thanks to the following for their contributions to this effort:

Partner Institutions: Cornell University, University of Florida, Washington University in St. Louis School of Medicine, Ponce School of Medicine, Weill Cornell Medical College, and The Scripps Research Institute

Indiana University collaborators: Digital Library Program, IU Libraries, University Information Technology Services, Data to Insight Center, School of Library and Information Science

VIVO
Enabling National Networking of Scientists

References

RIS2N3 *https://github.com/dgcliff/RIS2N3*
SEAD VIVO *http://vivo.sead-data.net/*
Sci2 Tool *http://sci2.cns.iu.edu/*
Apache Jena *https://jena.apache.org/*
SEAD project *http://sead-data.net/*
VIVO@IU *http://vivo.iu.edu/*
IU DLP *http://www.dlib.indiana.edu/*
IU Libraries *http://www.libraries.iub.edu/*