Ingest tools: home brew or off the shelf?

Major options

Data ingest for VIVO is a process of transforming an existing or new source of data into RDF and loading that RDF into VIVO's data store, called a triple store after the three-part data statements it contains. VIVO ships configured to use an open source triple store called Jena SDB implemented using one of several off-the-shelf databases – in most cases MySQL, although at least one VIVO site, Melbourne's Find an Expert site, is exploring IBM triple store technology implemented via Oracle. In addition, developers at Cornell and Florida are leveraging the new RDF API implementation in VIVO 1.5 to experiment with other triple stores including Sesame, Virtuoso, and Owlim.

Imagine first that you have a magic black box that converts each data source on your list into RDF compatible with the VIVO ontology. Loading that RDF into VIVO can be accomplished as simply as logging into VIVO as a site administrator and loading the RDF via the **Add/Remove RDF Data** command in the Advanced Data Tools section of the site admin menu.

Caveats

Unfortunately it's rarely quite that simple.

- The magic black box doesn't yet exist, although the tools to work with RDF are improving all the time and the VIVO ontology has gained traction as a standard for information exchange on research networking.
- Furthermore, unless you are starting with an empty VIVO, the process preparing RDF for VIVO will have to have be able to query your VIVO to make sure it's not duplicating data already in VIVO, including people, organizations, and the content of the dataset at hand. When your data comes to you from several sources, alignment based on names alone is prone to errors including false positive matches and false negatives, leading to duplicate URIs for the same person, organization, or other entity.
- Finally, the data you add will very likely not remain static. Your data ingest methodology very quickly must also serve as a data updating and data removal methodology.

Okay, you've heard that before

You likely have some experience with ETL (extract, transform, and load) processes and you've heard about these problems before. This is good – you are aware that while VIVO is the challenge you are taking on now, getting data into VIVO is not that different from other platforms.

Unmasking the black box

You have at least three choices:

- You can enter sample data into VIVO through its editing interfaces, export the data, and write your own scripts to produce data matching what you see. This sounds like a cop-out on the part of the VIVO community, but some people with a lot of ETL experience prefer to leverage tools they already know to produce a given target
    - The VIVO ontology team has developed a number of visual diagrams of the VIVO ontology at both overview and specific levels to help understand VIVO data, and may allow you to bypass or minimize time spent on sample data entry
    - The **Karma** tool from USC's Information Sciences Institute has been extended to support the creation of VIVO-compatible RDF.
    - Furthermore, there are an increasing number of open-source libraries for writing RDF with PHP, Java, and Python.
    - There are also commercially developed and supported tools including TopBraid Composer
    - If you know of other open source or commercial tools, please add links to them here)
- You can use the **VIVO Harvester**, a framework for gathering data, transforming it, performing matching against data already in your VIVO instance, and adding that data directly to VIVO, bypassing the data tools in VIVO. There is a learning curve to the Harvester, but a number of VIVO technical teams use it extensively and have benefitted from shared experience in improving the Harvester framework on an ongoing basis
- You can write tools to transform your data to RDF using an ontology of your choosing or creation, import that RDF into VIVO, and use tools within VIVO to align data with existing data and transform data to the VIVO ontology, as described below under "Working with semantic rather than scripting tools."

The VIVO Harvester

The VIVO harvester can be configured for a wide variety of tasks.

- Configuration files can be adjusted to get data from different sources and in different forms.
- The Harvester is modular. Some sites use parts of the Harvester to accomplish parts of their ingest, and use home-brewed tools for the rest.
- Sometimes a home-brew perl script can be more easily tailored to your special needs.
- A wide variety of tools are out there to combine with the Harvester.

The Harvester has been extensively documented throughout it's lifetime by its original developers at the University of Florida and through the work of other VIVO developers and implementers at other institutions. Please see the Ingesting and maintaining data section for full details.

Working with semantic rather than scripting tools

Ideally each significant source of data at an implementing institution will first be represented by its own local ontology that represents the data source as it is made available to the project. By reflecting the data as it comes to you in an ontology, you are in a better position to detect changes (either additions or deletions) in the source over time and can reduce or transform the data transferred to your local VIVO instance.

One of the strengths of semantic approach is that by creating mappings from that source ontology to the VIVO ontology, much of the work of processing the data is not only clearer but can then be accomplished without writing programming scripts or Java code. A programming approach might come more naturally to you at first, but may prove be more work and less transparent to maintain. Limiting yourself to simple data formats such as spreadsheets or . csv files can be the equivalent of using a very small pipe to connect the semantically-rich data in your source with the semantically rich data in VIVO.

Working in the RDF world as early as possible in the data ingest process will also train you for using tools available for querying data in VIVO itself (e.g., using SPARQL to run reports), making VIVO data available as web services for consumption on other websites, or for mapping data exported from VIVO into other tools such as the Digital Vita tool developed at Pittsburgh.

The logic and application of semantic mappings are discussed extensively in the recommended book, "Semantic Web for the Working Ontologist", including many short examples and a step-by-step introduction of RDF and OWL capabilities.

---

next topic: Typical ingest processes