VIVO Data - what and from where

Introduction

You've looked at VIVO, you've seen VIVO in action at other universities or organizations, you've downloaded and installed the code. What next? How do you get information about **your** institution into **your** VIVO?

The answer may be different everywhere – it depends on a number of factors.

- How big is your organization? Some smaller ones have implemented VIVO only through interactive editing – they enter every person, publication, organizational unit, grant, and event they wish to show up, and the keep up with changes "manually" as well. This approach works well for organizations with under 100 people or so, especially if you have staff or student employees who are good at data entry and enjoy learning more about the people and the research. There's something of an inverse correlation with age – students can be blazingly fast with data entry, employing multiple windows and copying and pasting content. The site takes shape before your eyes and it's easy to measure progress and, after a bit of practice, predict how long the process will take.
    - This approach may also be a good way to develop a working prototype with local data to use in making your case for a full-scale effort. The process of data entry is tedious but a very good way to learn the structure inherent in VIVO.
    - We recommend that people new to RDF and ontologies enter representative sample data by hand and then export it in one of the more readable RDF formats such as n3, n-triples, or turtle. This is an excellent way to compare what you see on the screen with the data VIVO will actually produce – and when you know your target, it's easier to decide how best to develop a more automated ingest process.
- The interactive approach, or the manual data entry, will obviously not work with big institutions or where staff time or a ready pool of student editors is not available. There are also many advantages to developing more automated means of ingest and updating, including data consistency and the ability to replace data quickly and on a predictable timetable. Some institutions have opted for utilizing the Karma data integration tool for producing RDF data out of the tabular data that comes from relational databases by modeling it in an interactive environment. Karma data integration tool has one advantage since its interactive visual environment helps in understanding how the ontologies work.
- What are your available data sources? Some organizations have made good institutional data a priority, and others struggle with legacy systems lacking consistent identifiers or common definitions for important categorizations such as distinct types of units or employment positions. It is very important that data you receive from legacy systems be examined and identifiers and names for people and organizational units are standardised and made consistent across the various systems. Those legacy systems may use different identifiers/codes and names for the same organizational unit and you want to ensure that data is clean before you start modeling it to the ontology. Another aspect of legacy data is that you may have to make some inquiries to find the right people to contact to find out what data sources are available, and the stakeholders on your VIVO project may need to request access to that data.

Next – what is different about data in VIVO?

As we've described, it's well worth learning the VIVO editing environment and creating sample data even if you know you will require an automated approach to data ingest and update.

VIVO makes certain assumptions about data based largely on the types of data, relationships, and attributes described in the VIVO ontology. These assumptions do not always follow traditional row and column data models, primarily because the application almost always allows for arbitrarily repeating values rather than holding strictly to a fixed number of values per record. Publications may most frequently have fewer than five authors, but in some fields such as experimental physics it's common to see hundreds of authors – not very workable in a one-row-per-publication, one-column-per-author spreadsheet model.

In VIVO, data about people, organizations, events, courses, places, dates, grants, and everything else are stored in one very simple, three-part structure – the RDF statement. A statement, or triple, has a subject (any entity), a predicate or property, and an object that can be either another related entity or a simple data value such as a number, text string, or date. While users will see VIVO data expressed in larger aggregations as web pages, internally VIVO is storing its data as RDF statements or triples.

This is not the place to explain everything about RDF – there are many good tutorials available and other sections of this wiki explain the VIVO ontology and the more technical aspects of RDF. For now, just bear in mind that while the data you receive may come to you in one format, much of the work of data ingest involves decomposing that data into simple statements that will then be re-assembled by the VIVO application, guided by the ontology, into a coherent web page or a packet of Linked Open Data.

What data can VIVO accept?

With VIVO, your destination will be RDF but you may receive the data in a variety of formats. A first stage in planning ingest involves analyzing what data you have access to and mapping on paper how it need to be transformed for VIVO.

It's probably most common for data to be provided in spreadsheet format, which can be very simple to transform into RDF if each column of every row refers to attributes of the same entity, usually identified by a record identifier. The process becomes more complicated if different cells in the same row of the spreadsheet refer to different entities.

The following spreadsheet would be very easy to load into a VIVO describing cartoon characters:

id	name	height	age
1	Goofy	89 cm	11
2	Elmer Fudd	60 cm	45
3	Roadrunner	140 cm	2

You can readily imagine a storing the information about each cartoon character – id, name, height, and age – in one entity for each character.

A spreadsheet of books, however, would be more complicated:

id	title	publication date	author	publisher	pages
497531	Cartoon Animation	1967	Wilcox, George	HB Press	237
501378	Animation Techniques	1989	Smith, Charlotte and Wilcox, George	Cinema Press	359
391783	Digital Animation	2005	Ivar, Samuel	Digital Logic, Inc.	327
34682	Dairy Barn Automation	2011	Wilcox, G.P.	University of Minnesota Press	403

VIVO stores the book, each author, and the publisher as independent entities related to the other. This enables information about the book, authors, and publisher to be queried and displayed independently, a key feature of the semantic data model.

We have also introduced a common problems with spreadsheets – when a cell contains more than one value. We need a way to connect the book, "Animation Techniques," with two authors, and to indicate that Charlotte Smith is the first author and George Wilcox the second.

This example also points out another challenge in working with data – it's not always clear when values that appear similar actually represent the same entity, whether a person, organization, title, journal, or event. It would be easy to assume the George Wilcox in the first entry is the same as G.P. Wilcox in the 4th, but they are writing about very different topics. For a small organization, it may be easy to disambiguate authors, but this becomes a major challenge at the scale of a major research university.

Data cleanup and disambiguation are challenges for any system and will be a common theme in documenting VIVO data ingest along with semantic data modeling that is more specific to working with VIVO.

Further topics

- Policy and planning questions for VIVO data
- Typical sources of VIVO data
- Public vs. private data
- Data source specifications for implementation
- Ingesting and maintaining data
- Manual Data Entry
- Managing authorship information during ingest and after in VIVO
- Using the Karma data integration tool
- How to manage data cleanup in VIVO

See also

Under Ingesting and maintaining data

- How to plan data ingest for VIVO
- Ingest tools: home brew or off the shelf?
- Typical ingest processes
- Challenges for data ingest
- Monitoring for quality

Under Maintaining VIVO

- VIVO Data Management