Starting a VIVO Project

There is perhaps no "typical" VIVO, since every organization from the smallest to the largest is different. Certain themes recur, however, in planning and executing an evaluation, pilot, and eventual full rollout of VIVO.

Finding data

In a very small organization such as a department, a single research center, or small institute, it may be possible to populate a VIVO instance interactively, either by training a few people to enter data about everyone or asking people to fill in their own profile information. This approach does not scale to large organizations, where much of the information desired for inclusion in VIVO already exists in institutional systems of record such as human resource databases, grants management systems, and course catalogs. Manual entry is time consuming, and if the data exist in a machine-readable form it makes little sense to re-key it.

When data systems already exist, one of the first challenges will be to find common identifiers to link them. Most organizations have evolved an identifier for people that can be shared publicly, whether the institutional email address or another number other than a social security number in the U.S. or other confidential government ID. It's remarkably hard to identify who should be in a VIVO, and surprisingly few large organizations even have an accurate or consistent scheme for identifying departments and organizations, especially somewhat transient units such as research centers.

Gaining access to data and permission to reuse it

The mere fact that data exist does not assure that they will be made available to you, however. It will be important to make contact with the owners or stewards of the data desired for VIVO, and it may not be trivial for them to make data available. It's common for public and private data to be intermingled in systems of record that have been built for functions such as payroll, performance evaluation, or financial accountability. Gaining access to data and permission to display it on the Web often requires building bridges to other units and catching the ear of people with the authority to support your requests. Some of the most successful VIVOs involve closely-coordinated collaborations among central administrators, IT staff, the library, and the research administration office.

Mapping data into VIVO

One of the challenges for the VIVO community is that while many organizations have similar kinds of data, there are a myriad ways in which employment and affiliation, grants, publications, courses, facilities, and other useful data for VIVO are stored and made available for reuse. Very often the data need some enhancement and structuring in order to work with VIVO's very granular approach to storing data.

Three forces are at work to make the task harder than it may at first seem. First, we've already mentioned the challenge of finding reliable and unique identifiers on people, organizations, courses, journals, grants, places, and even potentially events. Many systems rely solely on the alignment of text strings to associate one data point with another, and if (as is common) a person or department may be recorded with several name variants in different systems or at different times, VIVO will then have only partial information for each variant and the duplication of names will become very evident. Identifiers help to prevent data mis-attribution and alignment problems.

Secondly, VIVO's ontology may at first seem to align with local naming conventions, as for example with listing different publications types. Sometimes this can be handled by careful alignment, while in other cases it may be worthwhile to make what's called a local extension to the ontology. When adding a more specific sub-type of an existing ontology class, local values will then harmlessly roll up to the more general categorization when data from different VIVOs are combined.

Finally, VIVO may just be more complicated than single-purpose systems. For purposes of displaying a person's CV or a list of publications on their home page, storing publication citations as formatted text in a big block may be perfectly adequate. When building a database of publications, however, the fact that one person often co-authors with others at the same institution cannot be escaped, and a system that does not break out authorships and journals or publishers separately will accumulate redundant data and reduce the ability to generate reports or reuse data in different contexts with any confidence.

Data do not remain static

It's easy to focus on a one-time data load, especially in a proof of concept or pilot project. Putting up any kind of information system, especially one with as many different kinds of data as VIVO, requires making a plan for handling updates — both full and partial data removals as well as additions. Many data sources don't maintain the date of last modification so a complete scan is needed to detect change. In some cases data layers can be replaced wholesale in VIVO because they interact or interleave very little with other datasets, which simplifies updating.

Users will quickly point out incorrect or outdated information, and once the confidence of a user has been lost in the accuracy or completeness of an online database it's not easy to gain that confidence back. Some VIVO installations have augmented display pages with information icons that clarify the source of each category of information and even identify who to contact to change it. Where policies permit users to edit their own profiles, change is straightforward, but when data have come from a system of record it's more sustainable to make changes back at the source so that errors don't propagate back to VIVO or to other systems.

Be realistic

If you are fortunate enough to have buy-in from top management and a commitment for the staffing and IT resources to embark on a multi-faceted exploration and implementation of VIVO, more power to you. In many cases people interested in VIVO need to start small and build a case for VIVO based on a targeted need identified for a small segment of the organization, whether a campus, a government department, or a distributed organization.
Starting small and not taking on too many different kinds of data allows you to familiarize yourself with what are often new concepts and new tools. Data are often dirtier than advertised to be, and more will be left to you to clean up in the process of importing into VIVO. If you need to hire technical help, it may take longer than anticipated to find a qualified person familiar with Semantic Web technologies than for routine website development.

**Reaching out to engage your peers**

Creating and maintaining a VIVO is a bigger task than one person or one organization can typically handle on its own. As you gain experience, share what you have done with your peers, and invite their feedback and suggestions. We encourage participation in the weekly implementation calls, and they can provide a way to share innovations and get advice on how to move past obstacles or scale up to the next level.

For more information

- Planning a VIVO Implementation