# Data source specifications for implementation

## Data source specifications for implementation sites

### Starting points - typical data sources

The focus will be on identifying appropriate sources for data at each institution and providing that data in a format that can be loaded into VIVO straightforwardly without a lot of custom development. This approach will provide key data for testing and demonstration while also exposing important nuances that must be accommodated for a more customized and auto mated process of updating and removing data.

**General principles**

*avoid information with any privacy concerns – you probably don't need it. At Cornell, we have no information on age, sex, race, national origin, citizenship, leave status, termination dates, home phone number, and certainly not Social Security numbers.
*don't load what you can't maintain – it's better to have less information and have it be correct than information that is out of date. For example, Cornell's VIVO doesn't store any contact information but links to the Cornell directory.
*there are useful "lessons learned" points on content, ownership, buy-in etc. extracted from a blog post on NASA's social networking internet.

# VIVO data ingest conceptual model

### 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. The ontology should be very simple, but should reflect the structure of the data – e.g., at Cornell people are not linked to departments but hold jobs that are instances of positions which are assigned to departments. 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. In the HR example, you probably don't want to include the job and position information in VIVO, but just link the person to a department through an appropriate affiliation relationship (object property).

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 will be accomplished without writing programming scripts or Java code. A programming approach might come more naturally to you at first, but will be more work and less transparent to maintain. Stripping local source information down to bare essentials for insert into VIVO would be analogous to using a 1/4" 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.

### Public vs. private data

Whenever possible, leave non-public information out of the public VIVO, since including private information will complicate a user's picture of his or her VIVO profile and make the entire project more difficult to manage. Semantic web tools have been developed to share data by exposing it for direct consumption in other tools as well as for human eyes to read, and while the Vitro software underlying VIVO offers ways to limit the visibility of the data on websites, a complete RDF export of a VIVO database will be directly readable by other tools that may make no attempt to filter by any criteria.

As with any data and any software, this is a common sense balance of benefits and risks. There are many reasons to include data such as department identifiers in VIVO that should be hidden from view to avoid clutter but are essential for aligning new data; the project will add more ways for users to limit the visibility of certain research-related information a person may not wish to share, such as a network of informal colleagues or a new area of investigation. However, we see little to gain and much to lose by putting any confidential data into VIVO, such as salary history, termination dates, leave status, or identification information (age, sex, race, nationality, marital status, home phone or address, etc). Cornell's VIVO instance links users to the campus directory rather than holding contact information directly, since our HR system frequently lags employees' own updates of their contact information.

Should VIVO become a **System of Record** (SOR) at your institution? That's really up to you, but you need to carefully consider the risks as well as benefits. VIVO may well become the SOR for information such as research areas and keywords, brief statements of research purpose, and perhaps publications. For other information such as grants and appointments that are currently maintained elsewhere for administrative purposes, VIVO should remain a downstream consumer of SORs rather than seeking to supplant core systems. At Cornell the college administrators feel a pressing need to have a data mart that combines **all** the information they need about faculty, including HR, grants, courses, course evaluations, and assorted other information including some they track directly. They have wide-ranging requirements for running reports on that data, however, and need to include salary history, grades, performance reviews, and other data that would be much better managed through a data warehousing and report generation tool behind appropriate firewalls than by a VIVO instance designed for public information discovery and sharing.

### Top down and bottom up

The ontology and controlled vocabularies team at Indiana University are already defining the major components of the VIVOweb ontology, based on the evolution of the VIVO ontology at Cornell but aligning it more closely with modular ontologies already in wide use on the Web. Think of this top level ontology as a tree with several primary branches.

This approach does not conflict with a bottom-up approach to define local ontologies, especially as a way to model locally variable data in a logical language and use semantic tools to figure out how to map up these more detailed branches of the tree into the structure of the common ontology. The greater detail available locally can be maintained where desired, but the higher-level, more general view of the data in the common VIVOweb ontology will enable data from multiple sites to be combined without additional mapping or transformation.

There are two possible mechanisms being explored by the development teams for maintaining synchronization between the VIVOweb high-level ontology and local institutional VIVO instances. The first approach will likely use the same Subversion repository used for Vitro source code for the ontology, and modify Vitro to load its startup ontology from the latest ontology checked out from Subversion at any local site. A second alternative would have satellite vivo systems talk to a master system for the top-level ontology and/or push changes out from a central instance to distributed sites. Both methods would separate the ontology into editable and non-editable components to make it clear at distributed institutions which classes and/or properties are required for direct data sharing and indexing across multiple sites.

OWL and RDF work well to allow adding detail through sub-classing and sub-properties. The model of maintaining a high-level conceptual agreement will also allow sites that have conflicting modeling needs to operate independently, as long as local more detailed models can be mapped to a more general common model. One example of this need is likely to come in differentiating types of faculty – for some institutions there may be no meaningful distinction between teaching, research, and clinical faculty, while other institutions use entirely separate systems to manage different faculty tracks and would need to have those distinctions reflected in their local VIVO instance.

**See also**
*Why should we create VCard entities linked to Authorships?

# Provenance of VIVO-Cornell data

Information on a VIVO people page can be edited by the individual upon logging in using his/her Cornell net ID and password. However, profiles are generally at least partially populated via automated feeds from "authoritative" sources where possible, and via manual entry by student employees and librarian curators.

For visual overview of data feeds, click through the screen shots on attached slide detailing VIVO content sources.

## Overview

*Short bio: web page (manual entry) OR annual faculty reporting db (where available)

## Research info

*Research and scholarship focus: annual faculty reporting db (where available)
*Grants on which individual listed as PI, co-PI: Office of Sponsored Programs warehouse
*Research areas: annual faculty reporting db (where available) OR college strategic areas brochures/web pages (via manual entry)
*Impact statement: annual faculty reporting db (where available)

## Affiliations

*Name, title, primary departmental or job affiliation, departmental chairship and other admin responsibilities: HR(PeopleSoft db)
*Affiliations with other Cornell departments, fields, units: web pages (manual entry)

## Teaching

*Teaching focus: annual faculty reporting db (where available)
*Courses taught: System for Tracking Academic Records of Students (STARS course db (Oracle)

## Service

*Outreach focus: annual faculty reporting db (where available)

## Background

*educational background, professional background, awards and distinctions, news releases individual is featured in: annual faculty reporting db (where available), and web pages (manual entry)

## Publications

*listed publications, Cornell event speaker: manual entry (editor or self-editing)
*linked publications: automated from bibliographic dbs

# Top priority data

# Employment data

This includes people and their employment relationships with the institution, as may be distinct from secondary research or academic affiliations not likely as well represented or maintained in institutional systems of record.

## Scope

People and their employment relationships with the institution. Note that these data should come from institutional systems of record and not be editable by end users or even by most proxy editors.

### Data Types

*swrc:Employee (and possibly swrc:Graduate)
*swrc:Organization or its subclasses swrc:Department and swrc:University

### Basic properties of people

Data properties (text strings, numbers, dates)
*full name (swrc:name or rdfs:label, typically in form lastname, firstname intial)
*first name (swrc:firstName)
*middle name or initial (swrc:middle)
*last name (swrc:lastName)
*primary institutional email address (swrc:email)
*university (HR) title
*working title
*preferred title

Object properties
*swrc:worksIn | department or other unit
*may need to be treated as sub-property relationship distinguishing between academic and administrative appointments, at least for faculty.
*head of | department or other unit

### Basic properties of organizations

Data properties
*swrc:name or rdfs:label
*1 or more distinct identifiers (e.g, HR department code, sponsored research department code)

Object properties
*parent or child unit(s) | department or other unit
*has employee | employee

### Issues to decide at each institution

*which people (clinical as well as research? professional staff? administrative staff who participate in research management? how far down the academic food chain? graduate students?)
*how much of the organizational structure? (usually this is driven by the organizational structure of the HR system, which may differ from academic or research units. Secondary research center or academic program affiliations are not likely well represented or maintained in institutional systems of record, and informal lab groups likely have no formal institutional identity at all.)
*do you need contact information, or is that better as a link – the HR system contact information at Cornell often lags the LDAP system used for online directories, so we just link every person's VIVO profile to their Cornell directory listing
*which identifiers? Most institutions now have a unique employee identifier which is typically not published on web pages but is not considered sensitive data – avoid information with any privacy concerns – you probably don't need it.
*don't load data you can't maintain

## Planning for updates at the time of data acquisition

It's easy at first to focus on getting data into VIVO, but the harder task is to keep it updated. This means establishing a way to remove data from the system when employees leave the university, as well as being able to recognize when a person's job within the institution has changed.

This job is complicated by the fact that the institution very likely does not store employment information the same way that you will want to represent it in VIVO. At Cornell, for example, a Person occupies a Job that is an instance of a Position (some of which have multiple occupants), and only the Position is related to a Department. Each monthly update comes to us not as a list of changes but as the current list of employees; we must check current Job and Position data against an intermediate database holding the last download to determine when a Person is no longer an employee of a Department. All that information could be held in VIVO, but the extra Jobs and Positions would only clutter up VIVO.

We use an RDF database to store the latest Job and Position data, but this could be done in XML or a relational database, or the comparison could be done via CSV files.

# Grants data

## Scope

VIVO at Cornell has information from the sponsored research systems at Weill Cornell Medical College (a test load of about 200 records) and from the Cornell Ithaca campus.

Note, however, that at least at Cornell these databases only include sponsored research, and typically not contracts or projects funded by private corporations or foundations. The notion of a grant or contract may also include program-based internal grant funding opportunities. Since several Cornell colleges collect data on their faculty via an externally-hosted reporting system (Activity Insight, from Digital Measures), VIVO has to be able to represent this fuzzier notion of a contract or grant in faculty profiles while keeping this information (not from an institutional system of record) internally distinct. We also receive purely narrative information called "research in progress," that may involve no funding source at all.

Confusion can also arise between grant programs, on the awarding side (e.g., project or department or center-based internal grant programs or programs providing fellowships or travel awards to graduate students), vs. grants on the receiving side.

## Data types

*We have not yet assigned research investigators to be a different class, either through inference or direct assertion. Research grants were added to VIVO after people, and grant information is matched to investigators based on the Cornell public net id of the investigator. When an unknown investigator turns up in a grant import, an LDAP lookup is used to create a basic Academic Employee individual with that net id and the department affiliation(s) indicated in LDAP.
*the grant itself.
*the grant funding agency – membership in this class should be inferred by the existence of "provides funding to" property relationships to other grant funding agencies (there is a hierarchy) or "funds award" property relationships to grants, with individuals only directly asserted to be an Organization or subclass such as University or Government Agency.

Related types:
*grant program and subclasses such as demonstration grant program, training grant program.
*a grant, contract, or research in progress reporting response (from faculty reporting, not the institutional SOR – hopefully other institutions won't have to deal with this).

## Properties of grants

Data properties
*grant title
*institutional award id
*would be very helpful to have the award id from the sponsor to enable matching against NSF, NIH, or other agency data to the extent available; note, however, that many grants are awarded as subcontracts through another university or organization, and the SOR may not include the grant id from the root source of funding.
*start date and end date – these are not necessarily easy to determine and may change over time, especially with multi-year awards.
*grant status - the key used at Cornell to indicate.
*(optional) award amount – we have never included this in VIVO at Cornell.

Object properties
*award administered by : department (linked via a sponsored programs identifier on the department. Note that at least at Cornell the list of departments for HR is not the same as the list of departments for OSP, and they have different identifiers. OSP data also go back in time and must continue to represent units no longer in existence).
*has investigator : academic employee.
*has co-investigator : academic employee.
*funded by : organization.
*subcontracted through : organization

Other useful data not likely available from SOR
*grant abstract (not available from SOR at Cornell)
*grant keywords (ditto)
*a property or properties linking grants with research facilities they either use or have provided funding to establish, but the data are not likely available unless compiled by the facility staff.
*if faculty are required to produce impact statements (maybe more a Land Grant notion), a link from the grant to the impact statement.

## Planning for updates

The institutional grant identifier should be a reliable way to detect grants that have already been added to your VIVO instance via a previous data load.

The grant status seems to be a safer way to determine when a grant has terminated than relying on 1 or more of the several data fields with the OSP database. Whether to remove grants that have terminated is an important policy question that touches on the issue of whether VIVO will represent a snapshot in time for current awareness purposes or is expected to remain a repository for information that rapidly goes out of date.

## VIVO is not an historical archive

Be very cautious in promising to maintain an historic record of grant activity (or any other information) – faculty will typically not want old information shown (or only very selectively), it may confuse users, and we don't plan to organize the VIVOweb ontology or display functions around hiding old data.

We have learned the hard way at Cornell about the distinction between a reporting system, for which time windows and comparisons between what was reported this year vs. last year are critical, and a current information discovery system. Because VIVO is capable of integrating data from multiple institutional sources in a much more convenient way for administrators to access, they sometimes push to change it to meet their reporting needs, which are quite distinct from those of the researcher or the casual user of VIVO.

## Publications Overview for the VIVO project

While the primary source of publication data for the VIVO project will be PubMed, it is important to also search other databases with life sciences-rich content such as the Web of Knowledge or SCOPUS. Both dbs also do a good job of providing access to social sciences literature, which would acknowledge the increasingly inter-disciplinary nature of biological research. PubMed has an advantage over WOK or SCOPUS in that the XML data produced from the PubMed web service is a logical, uniform source of publication information that will work across all the implementation sites.

We will be using the Bibontology http://bibliontology.com/ with a few extensions. For more about publications-related ontologies for VIVO, see the Ontologies Relevant to VIVO and ontology work of interest pages in this wiki.

# Other potential additions

## Additional people data

Bio sketches, affiliations, educational background, research interests.

## Research areas and keywords

### National and international sources

*URIs for MeSH terms
With regard to URIs for vocabulary entities, I have already completed this task for the NLM MeSH 2010 thesaurus (see attached abstract for AMIA CRI 2010), and will be adding additional major vocabularies including UMLS in the coming months.

For examples of URIs/URLs for MeSH2010 consider the following examples from the PDS MeSH2010 web service:

http://pds.portaldoors.net/mesh2010/d000001

for a single Descriptor Record embedded in PDS Message wrapper, or

http://pds.portaldoors.net/mesh2010/resrep/search?nam=informatics

for a search that returns a set of records each with its own URI. A more detailed paper on the PDS MeSH 2010 web service is in preparation.

Note that the general PDS framework allows for flexibility in distinguishing URIs from URLs so that any entity identified by a URI may have multiple locations that may include mirrored sites and/or different kinds of access. However, in the case of the current PDS implementation of the MeSH2010 vocabulary, the convention is adopted that the URL is the same as the URI so that it can be accessed readily from the web service which does not yet have any mirrors.

Also, please note that I do include the statement as required by NLM about NLM being the creator and maintainer of the original source Descriptor Records. This statement appears in every response returned by the PDS MeSH2010 web service.

Should anybody wish to discuss, I hope that you will visit with me in person at AMIA CRI 2010 in San Francisco next month.

### Local sources

## Facilities data