VIVO

Connecting People, Creating a Virtual Life Sciences Community

[Medha Devare](#), [Jon Corson-Rikert](#), [Brian Caruso](#), [Brian Lowe](#), [Kathy Chiang](#), and [Janet McCue](#)
Albert R. Mann Library, Cornell University, Ithaca, NY 14853

Points of contact for this article:
Medha Devare, Bioinformatics and Life Sciences Specialist, Services and Collections, <mhd6@cornell.edu>
Jon Corson-Rikert, Head, Information Technology, <jc55@cornell.edu>

Abstract

To help faculty, researchers, and students in the life sciences discover common interests and make connections, the Cornell University Library has created a virtual life sciences community that uses an entity-relationship ontology model to organize and present information on people, research, and education activities. This single point of access for scholarly activity in the life sciences at Cornell – VIVO (http://vivo.library.cornell.edu) – transcends campus, college and department structure to provide Cornell faculty, students, and administrative and service officials, prospective faculty and students, external sponsors, and the public an integrated view of the life sciences at Cornell. At the request of university administration, the VIVO database is currently being augmented to provide similar content for the social sciences, engineering, physical sciences, international activities, and potentially other areas at Cornell.

VIVO's search engine clusters results into categories ("People", "Activities", "Events", "Organizations", "Publications", etc.), providing a campus-wide search capability that offers a richer context than typical undifferentiated text-based indexes. Each page in VIVO presents bi-directional hyperlinks to any related entities, giving users access to a web of content that might otherwise necessitate discovering and visiting multiple websites. VIVO draws on content from across the university through automated ingest of core university human resource and grants information, subscriptions to licensed and public domain publications databases, department or college faculty reporting mechanisms, and manual curation. The underlying ontology model organizes information at a granular level by type ("entities") and relationship ("properties") to present, for example, faculty profiles with affiliations to departments, fields, or research units, research projects, courses, seminars, and facilities relevant to life scientists regardless of the campus, college, or department in which the entity resides.

Background

Many areas of academic research have been moving inexorably in the direction of interdisciplinary collaboration as Internet technologies have enabled academics to make connections with each other and find online research materials with increasing speed and accuracy. To remain competitive, universities today must have the ability to take an integrated view of traditional disciplines from across the intellectual spectrum, and to transform that view into academic programs calling on multiple disciplines.

In recognition of this new paradigm, large institutions such as Cornell, Stanford, Harvard, and the University of Minnesota – among others – are encouraging inter-disciplinary collaboration among their faculties in a variety of ways. Cornell launched two multi-million dollar inter-disciplinary initiatives – the Genomics Initiative, now known as the New Life Sciences Initiative (NLSI) and the Social Sciences Initiative in 1997 and 2004, respectively, to recruit competitive faculty and students for new positions through a "department open" faculty hiring policy spanning traditional departmental boundaries. In 1998 Stanford began an interdisciplinary initiative called Bio-X, designed to strengthen collaboration amongst medical, engineering, chemical, physical and biological researchers

(http://biox.stanford.edu/about/history.html). Harvard recently released a study entitled "Enhancing Science and Engineering at Harvard," commissioned to address the difficulties in breaking down barriers to inter-disciplinary collaboration[1]. The President's Inter-Disciplinary Initiative on Arts and Humanities was begun at the University of Minnesota in 2004, and among other outcomes mandates the hiring of leading faculty dedicated to interdisciplinary scholarship in the arts and humanities[2].

One of the most frequent recommendations of such initiatives is fostering greater *interaction*, with the goal of catalyzing networks of campus-wide scholarship, research and educational activities. For such an inter-disciplinary scholarship and research model to truly succeed, rapid and well-organized access to information on scholars and researchers across disciplinary, structural, and often, geographic boundaries is necessary. For instance, interdisciplinary collaboration and resource-sharing as mandated by the NLSI at Cornell presupposes an awareness of potential collaborators not only across a few departments within a college, but across departments, institutes, and research centers within almost each of the University's 14 colleges, which reside on 4 geographically distant campuses. The desired outcome is best articulated on the NLSI website as being the continued drawing together of "biologists, physicists, chemists, computational scientists, and engineers in an atmosphere where traditional departmental and college boundaries become secondary to the intellectual work itself" (http://www.lifesciences.cornell.edu).

It rapidly became clear to the representatives from Cornell's Albert R. Mann Library who attended NLSI planning meetings that the researchers were severely handicapped in fulfilling the mandates of the NLSI by Cornell's large size and hierarchical view of colleges, departments, and other units, replicated in the University's online presence. Consequently, scientists – as well as those attempting to showcase or publicize their research (e.g., the News Service) or attract donors (e.g., the Office of Alumni Affairs and Development) – found themselves in poorly charted terrain, facing problems in multiple arenas, including:

- Devising an integrated presentation of "the life sciences at Cornell" to the Cornell community and beyond (the public, prospective donors, etc.)
- Navigating from one relevant "data point" – e.g., a faculty member's website – to similar activities elsewhere at Cornell
- Identifying appropriate collaborators/area experts, courses, events, and research technologies and facilities within and beyond their domain
- Following recent Cornell research activities and publications, especially by discipline rather than departmental affiliation
- Recruiting faculty and students: potential candidates were unable to easily penetrate beyond single-college or department websites to obtain an accurate representation of the breadth and depth of life sciences faculty and facilities at Cornell

The Cornell University Library recognized in early 2003 that as a neutral entity with a clear mandate and strengths in the curation, management, and dissemination of information – both print and digital – it could alleviate many of these problems. To that end, the Life Sciences Working Group, a multi-disciplinary group of science librarians with expertise in fields ranging from chemistry to veterinary and human medicine, was created. The group envisioned a life sciences portal to serve the needs of students and faculty at Cornell, both to learn of educational and research activities at Cornell and to discover high-quality resources outside Cornell.

To help researchers, students, entrepreneurs, and donors make connections with each other, the Life Sciences Working Group created a new portal – VIVO (http://vivo.library.cornell.edu) – based on an ontology model that transcends campus, college and department structure to provide an integrated view of the life sciences at Cornell. The primary goals of the portal are to offer Cornell faculty, students, and administrative and service officials, prospective faculty and students, donors, and the public a wide-ranging perspective on multiple facets of the life sciences within all of Cornell's colleges.

VIVO includes content related to any aspect of biology or the life sciences – broadly defined – at Cornell, and provides:

- A means for current or prospective students to discover fields of study, research projects, advisors, and opportunities
- A resource for researchers at Cornell or elsewhere looking beyond their immediate field for activities or collaborators relevant to new explorations

- A source for industrial partners and donors to evaluate the depth and breadth of Cornell's involvement in areas of common interest, and seeking points of contact
- A resource for administrators or journalists seeking information and expert commentary on current hot-button issues or more general scientific background
- A source for outside educators, both to inform themselves and to access lesson plans, course websites, presentations, or audio and visual materials

These needs no doubt exist at other institutions moving towards inter-disciplinary scholarship. Some of the recommendations of the "Enhancing Science and Engineering at Harvard" study[1] released by Harvard University in 2006 articulate the necessity for:

> "**University-wide information databases**: Harvard's scale and dispersed geography make it difficult to know what other research is taking place across the University or what lab and research capabilities exist in other departments, schools or hospitals. **No searchable databases exist for research or teaching expertise, and there is no central repository of Harvard faculty abstracts and ongoing research.**"

> **Exploiting technology to shrink distances and build linkages**:...We encourage the University to use technology to facilitate interactions between campuses and make it easier for faculty members to be aware of research being conducted elsewhere in the University. **Technology solutions can range from a searchable database of Harvard lab and research capabilities to a searchable repository of abstracts of Harvard research proposals.**

> **Address shortcomings in Harvard's research and instructional technical infrastructure**:...Coordinated acquisition and management of costly facilities, such as MRI machines and computational resources, would yield big dividends. In addition, Harvard should invest in technical infrastructure for innovative teaching. We propose three steps toward meeting these goals: **Establish a database of the interests and capabilities of Harvard's scientists and engineers to increase connections among them**..."

At Cornell, the librarians involved in creating VIVO employed a "build it and they will come" approach after finding it easier to demonstrate than to describe the benefits of a unified discovery tool. Early engagement of key stakeholders including Cornell administrators and new life sciences faculty generated positive feedback and requests for additional coverage as well as the ability to support selective filtering.

A portal focused on research within the College of Agriculture and Life Sciences and hosted alongside VIVO was released in 2005 (http://research.cals.cornell.edu/). Another portal followed in 2006, highlighting entrepreneurship instruction and activities across Cornell and leveraging VIVO content through web services. The Office of the Provost has now also committed two years of funding to extend the underlying database into additional disciplines to include all faculty and scholar research profiles across the campus. The Library views this effort as a fundamental enabling infrastructure for the truly inter-disciplinary university, with the academic lives of scholars and researchers depicted in all their richness (departmental affiliations, grants, courses taught, publications, research descriptions, and more).

Functionality of VIVO

Normally the most efficient strategy for delivering large amounts of content via the Web is to structure the content to match the plan for delivery on the website – ensuring that the desired browsing and searching behaviors are built into the site structure at the most fundamental level. However, when data are stored as undifferentiated text fields in databases optimized for a single mode of access, much of the potential for flexible delivery of content is lost. For example, if publications are stored as part and parcel of faculty profile database entries, it may be impractical to filter and display them by department, college, discipline, journal, or time window.

The time required to accumulate and update content, whether by faculty or on their behalf, must increasingly be justified by leveraging that content for multiple purposes on multiple websites, as well as for internal administrative reporting. To this end VIVO adopts a more atomic-level content model, using a uniform structure of entities and property relationships among them to allow the maximum flexibility for managing content and assembling it for delivery. What appears as a one-page view in VIVO may pull from several, a dozen, or even hundreds of separate standalone entities linked together via a defined web of cross-relationships.

This apparently contradictory notion – applying a simple, uniform structure to afford the greatest diversity of output – has strong roots in computer science dating back to the early frame system research by Marvin Minsky at MIT[3]. While not breaking new theoretical ground, VIVO achieves unusual fluidity in delivering information to the web by storing a minimum amount of information in entity records and using relationships to encode the many additional attributes that constitute most of the richness of VIVO's content.

VIVO also supports grouping entities in arbitrary or structured ways, most commonly to provide entry points for browsing content using a tabbed document interface. Tabs such as "People", "Academic Units", "Education and Training", and "Research Tools" on the home page shape the presentation of its content and indicate the scope of the site (Figure 1).

Figure 1. Home page of VIVO (http://vivo.library.cornell.edu)
(For a larger view, click here.)

The curators for the site create tabs themselves using an interactive interface that is based on filters to draw out appropriate entries by type, affiliation, or time period, and then arrange the tabs horizontally across the page or nested in hierarchies. Because VIVO includes labs, centers, departments, facilities, research-related resources, seminars, grants, and publications as entities in their own rights, curators have the flexibility of providing many more points of entry for browsing research-related activities than through faculty profiles alone.

The four currently populated top-level tabs in VIVO besides "Home", "People", "Education and Training", "Research Tools", and "Facilities" are described next.

People

Faculty, researchers, and administrative staff supporting the life sciences may be viewed alphabetically, through a gallery of images, or via searches. Unlike typical faculty web pages at Cornell, these profiles are not department-

centered, but acknowledge the diversity and richness of a person's activity and academic life by identifying all of a faculty member's department, field, and research center affiliations at Cornell. As shown in Figure 2, faculty pages use relationships that show the richness of their academic lives, including among others:

- grants for which they are principal or co-principal investigator
- Cornell seminars at which they may have spoken
- courses they teach, with links to all cross-listing departments so that the user may find information about the department or other courses offered by it
- recent articles downloaded every 90 days from the databases *Biosis* and *PubMed*

Figure 2. Example of a faculty profile page in VIVO.
(For a larger view, click here.)

Education and Training

Information on graduate fields of study, course listings, seminar series, literature resources, workshops, and a chronological list of seminars for the upcoming two weeks is available. Each of these includes information on the seminar speaker and his/her academic affiliation, and is associated with the URL of the host department's seminar series listing.

Research Tools

This section of VIVO was conceived as a tool box for faculty, staff, and student researchers in the life sciences. Content consists of public domain databases important to life scientists, organized from the micro (biomolecules, genomics and proteomics) to the macro scale (cells and tissues, organisms, populations). Each tool includes a sentence to help the user determine if the resource will be of use to him/her, along with a link to a longer description of the resource.

Facilities

With the life sciences at Cornell spread over about twelve colleges, several departments, and at least 3 campuses, scientists are often unaware of analytical facilities that may be more accessible and/or less expensive than others outside Cornell. The "Facilities" section of VIVO is an attempt to curate facilities by the type of technology (Genetic/Genomic/Molecular Biology; Greenhouses; Nanotechnology/Microfabrication, etc.) regardless of which department, research center, or college the facility resides in.

Content Creation and Management

VIVO has never been intended to replace the myriad research and public information websites relating to life sciences at Cornell (the total Cornell search space exceeds 10 million pages), but instead provides a discovery and navigation layer above these, allowing users to follow connections across Cornell based on common subject areas, affiliations of individual faculty and researchers, organization units, or resources independently of administrative boundaries.

Initial content curation for VIVO was accomplished largely by the authors and a student employee, with significant direction and feedback from members of the life science working group, the CALS Research Portal curators, and curators responsible for expanding the database into disciplines beyond the life sciences. Maintaining a site with such far-ranging content as VIVO contains is a challenge, yet it is equally clear that the value of a resource like VIVO owes much, if not everything, to being current and updated. It was clear from the start that manual curation alone would not suffice to maintain VIVO. Thus, content is now gathered via data feeds in a more or less automated manner, as well as via manual curation. This blend of manual and automated entry is a critical element in achieving sustainability for the site.

Automated retrieval of content

The development team very early sought out machine-readable sources of content for VIVO, beginning with grants from the university's Sponsored Programs Data Warehouse and journal article citations from Biosis and Pubmed (to be supplemented in 2007 with ISI Web of Knowledge). Additional data elements not available from Cornell or external databases have been solicited directly from faculty through the annual college faculty reporting process, including free-form keywords and association with identified collaborative research areas.

A basic entry for each person with title and department affiliation is created from core data extracted from the university's PeopleSoft human resources database, including academic and professional staff titles, contact information, and employment-related appointments. Comparisons from month to month pinpoint new arrivals and identify those people who have left Cornell through death, retirement, or resignation. Given the need to continue to reflect the activities of emeritus faculty, some human involvement is still required. Information on grants is also being managed with a clear preference for the authoritative source, limiting the ability for interactive editing to assure that VIVO matches the appropriate official information source. In order to avoid the challenges and risks of handling sensitive information, the VIVO database only holds publicly visible information about a person or grant. This approach has been greeted enthusiastically by the administrators of Cornell's data warehouses, who are eager to have their publicly visible data used and verified through greater exposure within Cornell. When errors are discovered, changes are made at the source rather than in VIVO.

A parallel but independent effort by VIVO technical staff to work with administrative and IT officers of Cornell's 11 colleges is helping to coordinate a more uniform approach to faculty self-reporting across Cornell. Administrators at Cornell have recognized that consistent and up-to-date information on faculty activities is critical to student and faculty recruitment, public communication of the university's activities, and academic vitality within the university – not to mention Cornell's current $4 billion development campaign. Devising a single approach to meeting both distributed and central information needs has been challenging, but a task force representing 9 colleges and the library has recently begun coordinating an approach to faculty reporting that addresses both internal administrative (and often sensitive) reporting and the need for improved content on faculty activities for display on the Web. With

the diversity of Cornell's subject areas of study, no single set of data elements (nor even agreement on who has faculty status) can easily be defined, and each college will continue to maintain an independent data warehouse to augment the information stored with outreach activities, clinical practice, industrial consulting, or course evaluations.

VIVO's database plays an important role in this faculty self-reporting effort by providing a path for colleges to share data without relinquishing control, to retain distinctions between clinical and teaching faculty in Veterinary Medicine, for example, while reporting faculty at an aggregated level university-wide. The VIVO team currently adapts ingest procedures to each local data structure, which can range from spreadsheets and Filemaker databases to complex relational models and XML. Through the current reporting initiative, many of the colleges at Cornell will share a reporting system, which will simplify data integration for VIVO in the future. As further incentive for sharing data with the VIVO database, the library has developed web services to deliver aggregations of content back to colleges and other units, presenting a focused view of research and related activities at Cornell. First examples include the entrepreneurship portal (http://entrepreneurship.cornell.edu) mentioned above, a faculty directory across all departments of the College of Engineering, and an international gateway highlighting research, extension activities, and faculty expertise by country and region.

Manual curation of content

Curators enter content into the database via forms that can be used with minimal training and no knowledge of programming. Each entity is classified on entry by type, i.e., a faculty member is identified via a pull-down menu on the entry form as the type "*Cornell faculty*". The URL for the person's primary departmental affiliation is entered along with anchor text, and a research description is compiled from their websites. URLs for additional departmental, field, or research unit affiliations can also be added. The faculty member's photograph can be saved onto the curator's computer and uploaded; a "citation" field allows the curator to acknowledge image copyright and/or enter a caption as appropriate. Time-sensitive entities such as seminars, news releases, and conferences or workshops can be sorted for display via date and time fields.

In keeping with the underlying ontology model described in more detail below, the type of an entity then determines what *properties* it has, or what relationships may be made to other entities and what types the related entities may be. After entering the basic information for the faculty member, a curator selects each desired property relationship and picks the appropriate related entity from people, departments, rooms, or seminar series according to the type of linked entity is defined for that property in the ontology model. If the related entity does not yet exist, the option of creating new entries is available.

For instance, a person's entry can reflect the fact that they are a member of a department via the relationship "*faculty member in*", which leads to an appropriate pull-down menu of Cornell departments (Figure 2). Primary investigator status on grants or authorship of publications are indicated via the relationships "*primary investigator of*" and "*author of*", respectively. Other relationships indicate a person as being "*featured in*" a news release or participating in strategic research areas ("*has research interest*") or conducting research in a particular geographic area of the world ("*has geographic research area*"), or being a "*speaker at*" an event.

These property relationships and their connected entities combine with data unique to each entity (in the case of a faculty member, this is typically their educational or professional background or research description) to create a seamless display – a profile page highlighting all the activities and affiliations of a person. What differentiates VIVO pages from the typical university faculty profile web page is not only that many diverse data elements are integrated into a single page, but that each element within the page is a link to the full context of the related information – a department profile, information on the funding source and co-principal investigators of a grant, or the complete citation, co-authors, and (if available) a link out to full text of a journal article. A user discovering a VIVO faculty profile via a search engine can navigate directly to the person's department to investigate other faculty or courses, to the parent seminar series of a seminar given to see what other presentations might coming up, or to a research area to view other participants, programs and resources.

The related entities in turn link back to the faculty profile and on further to other elements of related content. This seamless, bi-directional navigation most distinguishes VIVO from typical top-down, scripted navigation schemes, creating an interconnected web of content with unlimited entry points.

Content Discovery

The VIVO model is particularly well suited to today's information discovery patterns, where users frequently start with Google or another Internet-wide search engine to find people and activities even within their own institutions. Rather than navigating to the home page of a website, many users search first and arrive at any page within a site that has been indexed and ranked by a search engine based on factors that may bear little or no relationship to the rest of

the website. VIVO is optimized to support this unscripted pattern of website use since it provides context for each person or resource discovered and offers ready-made links to related entities. Rather than each page serving only as the end point of a search, it can become the launch point for further investigation, reducing the need for iterative searching while facilitating serendipitous discovery of similar content.

VIVO also has its own native search functionality powered by the Lucene open-source search engines. While Lucene supports ranking, VIVO's default search display categorizes results first by type rather than relevance ranking. For example, a search for "microarray*" shows results organized into categories such as people, activities, organizations, online resources, and publications, providing rapid access to exactly the sort of information sought (Figure 3). Each of these entries then links directly to the original resource, whether that is the full text of a recent article, the web page describing a faculty member's research, or a new genomics service. The important point to note is that the people appearing here are from departments at Weill Cornell Medical College, the Veterinary College, the College of Agriculture and Life Sciences, the College of Engineering, the New York State Agricultural Experiment Station, and the Boyce Thompson Institute, clearly illustrating the cross-disciplinary functionality of the database. Cornell's primary search page will undoubtedly return results as well, but because the context and the typing of results are missing, the user has to spend time scrolling through multiple pages to find the information of interest to him/her.

VIVO
virtual life sciences library

Search [microarray*] [Go]
Advanced Search | Search Tips

Cornell University Library

Home | People | Education & Training | Research Tools | Facilities   Index | About | Contact Us

Life Sciences in the Library | New Life Sciences Initiative | Cross-campus and Collaborative Initiatives

Cornell University Library » VIVO

Jump to results of type: people | activities | organizations | online resources | publications | technology

people

Cornell Faculty (21)

- Blissard, Gary W. | Scientist and BTI Vice President for Research | *Boyce Thompson Institute profile*
- Buckler, Edward S. | USDA-ARS Research Geneticist | *Buckler Lab web page*
- Coonrod, Scott C. | Assistant Professor | *Weill Cornell profile*
- Jaffrey, Samie R. | Assistant Professor of Pharmacology | *Weill Cornell profile*
- Jander, Georg | Assistant Scientist | *Boyce Thompson Institute profile*
- Lee, Kwangwon | Assistant Professor | *Plant Pathology profile* | *New Life Sciences Initiative profile*
- Lei, Xingen G. | Associate Professor | *The Lei Lab web page* | *Animal Sciences profile*
- Lin, David | Assistant Professor of Neurobiology | *Biomedical Sciences profile* | *New Life Sciences Initiative profile*
- Liou, Hsiou-Chi | Associate Professor | *Weill Cornell profile*
- Mezey, Jason | Assistant Professor | *Biological Statistics and Computational Biology profile* | *Mezey Lab web page*
- Nixon, Alan J. | Professor of Large Animal Surgery | *Clinical Sciences profile*
- Setter, Timothy L. | Professor | *Crop & Soil Sciences profile* | *Plant Breeding & Genetics profile* | *Plant Biology profile*
- Sinha, Animesh A. | Assistant Professor | *Weill Cornell profile*
- Smart, Christine D. | Assistant Professor | *NY State Agricultural Experiment Station profile*
- Tye, Bik-Kwoon | Professor of Molecular Biology | *Molecular Biology & Genetics profile*
- Wadler, Scott C. | The Richard T. Silver Distinguished Professor | *Weill Cornell profile*
- ⊞ 5 more

Cornell Academic Staff (1)

- Yu, Long-Xi | Research Associate

activities

Cornell Semester Course (1)

- Statistical Theories Applicable to Genomics | 4 credit course | *MATH 675*

Graduate Field/Program (1)

- Plant Biology | graduate field | *Graduate School Field web page* | *The Graduate Field of Plant Biology web page*

Research Grant (4)

- ACQUISITION OF INSTRUMENTATION FOR ESTABLISHING A MICROARRAY CORE | Research Grant
- FROM PROPLASTID TO CHLOROPLAST: UNDERSTANDING PLASTID DIFFERENTIATION IN MAIZE THROUGH MICROARRAY AND PROTEOME ANALYSIS | Research Grant
- LINEAR MAPPING AND MICROARRAY DEVELOPMENT | Research Grant
- MICROARRAY ANALYSIS OF CARDIAC GENE EXPRESSION IN GERMAN SHEPHERDS WITH SUDDEN DEATH | Research Grant

Research Related Service (3)

- Center for Gene Expression Profiling (CGEP) | cDNA microarrays | *CGEP*
- Institute for Computational Biomedicine (Weill Medical) – Microarray Resources | research consulting service | *ICB Microarray Consulting*
- Microarray Spotting, Analysis, and Consulting Service | cDNA microarrays | *Big Red Spots*

organizations

Academic Department (1)

- Plant Biology (BIOPL) | Cornell department | *BIOPL web page* | *Plant Biology courses*

Research Program, Unit, or Center (1)

- Max Planck Institute for Chemical Ecology Department of Molecular Ecology | research department | *Department web page*

Service Laboratory (4)

- Computational Genomics Core Facility | laboratory service facility | *CGCF*
- Life Sciences Core Laboratories Center (formerly Biotechnology Resource Center; BRC) | core service facility | *LSCLC web page*
- Microarray Core Facility | core service facility | *Big Red Spots*

Figure 3. Sample search in VIVO: microarray*.
(For a larger view, click here.)

Technical implementation

The Life Sciences Working Group knew the proposed portal would need to be flexible – it would not have been possible to draw a site map or design the interface in advance, nor did the group have an existing collection of documents or data to reference. Group members acknowledged that the content would start small, with a few predictable content types (people, departments, courses, publications) but extend organically based on the relationships each of these individual entries would have with other people, activities, and organizations.

What is the VIVO ontology? Entity-relationship models are becoming increasingly common as tools to organize and present knowledge. The term "ontology" originates in the field of philosophy as the study and characterization of what things exist in the universe. However, the fields of computer and information science have adopted the term to refer more specifically to models of entities and their relationships or interactions[5] or "a formal representation of technical concepts and their interrelations in a form that supports domain knowledge" (from Raskin & Pan, http://sweet.jpl.nasa.gov/). In the context of these definitions, VIVO would not be considered a life sciences ontology because it categorizes entries more by affiliations and associations than by scientific concepts and terminology.

VIVO stores information as instances of types linked by object relationships defined to connect explicit types – for example, between a faculty member and a course he or she teaches. Each entity has very limited internal data fields; however, each type of entity has a configurable number of labeled data properties, and inherits any data properties of its parent type, as shown below in Figure 4. These data properties are most frequently free-text format but may also be numeric or date information. When editing a member of any type, the data properties (if any) associated with that type and its parents are presented for entry or editing.

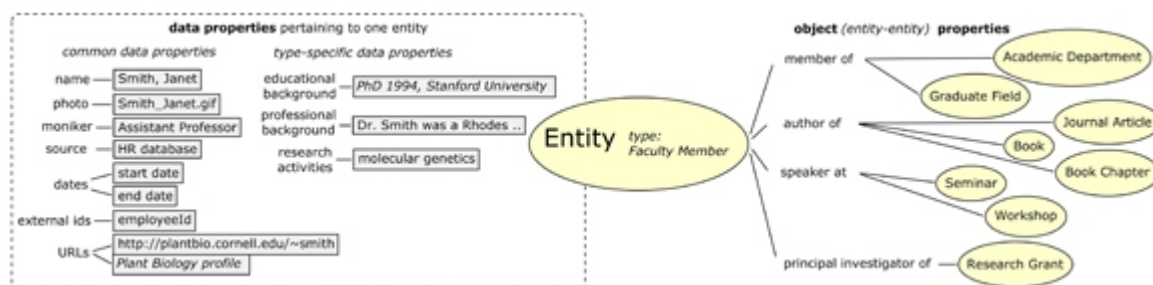**Figure 4. Sample entity structure for a faculty member showing internal common data properties as well as object property relationships with other types of entities.**

Data properties provide an additional layer of flexibility beyond the traditional entity-relationship (type/property) structure, allowing singular text values to be maintained in the database through simpler interfaces than those

available for adding new object entities. They greatly facilitate alignment of VIVO with external data sources through support for granular data elements and external system identifiers. Labels and categorizations of data rooted in external systems can be maintained in VIVO for data archiving, integrity checks, and updates – all critical factors for sustainability.

The structure of the types and properties in VIVO – the underlying ontology specifying what each entity is and what relationships it is defined to have with other entities – remains independent of the content that populates the database and is presented as the website. This allows VIVO to be cloned either as a ready-made application accepting similar content in different disciplines or as an empty framework ready for populating with types and properties that may be very different: from family or organizational histories to models for tracking agricultural extension and training programs in India.

Impact

VIVO's campus-wide, cross-referencing search capability and large index of life sciences researchers, resources, and facilities make it a core service whose timeliness and need at Cornell and beyond are becoming clear to faculty and administrators alike, as evidenced by some of the comments we have received:

> "...VIVO helps build a "virtual" research community...[which can] have several advantages – it is fluid, can change relatively rapidly and can present itself in more interesting ways to the rest of the world than more traditional structures."

> "As an administrator, it is often difficult to access information in a manner that is quick and in a form that is useful. VIVO and the CALS portal could be a major step forward in reducing the amount of time we spend collecting information and in better utilizing [it]."

> "Vivo provides unparalleled access to information about the life sciences at Cornell in a user-friendly way. This will be of particular benefit not only to those researchers and students already at Cornell, but potential faculty and students as well, by offering a much-needed, integrated view of the life sciences community at Cornell."

> "VIVO saved my life as a new faculty member at Cornell; I used it all the time to find facilities and people I might work with."

Rapid, well-organized access to the wide range of life sciences-related information across disciplinary, structural, and geographic boundaries had not been possible at Cornell until the development of this database. VIVO seeks to reach the widest possible audience by leveraging – not supplanting – existing online descriptive information to allow the materials prepared by information providers within the life sciences to be discovered via as many paths as possible.

VIVO is helping to change the culture of information technology at Cornell by harnessing data from centrally managed data warehouses and providing integrated but appropriately filterable views to distributed units. From the perspective of the central IT administration, VIVO affirms the value of their new service-oriented architecture (SOA) model. For the distributed units, VIVO delivers a rich, consistent feed of data never before available from a single source while also providing additional visibility on websites across the university. From the perspective of the library, VIVO fosters closer relationships with academic units and individual faculty while showcasing the library as an active, engaged partner that uses its information management capabilities in innovative ways to support the intellectual life of the university.

Providing cross-disciplinary views across universities is invaluable in this era of the need for increased collaboration among scholars and scientists. We anticipate that VIVO, and services like it, will only gain value by providing educational and research information in the life sciences, and catalyzing the formation of synergies and linkages to the mutual benefit of researchers, students, administrators, and donors. In short, VIVO embodies a virtual community, the creation of which provides a hitherto missing but necessary service that enables stakeholders to participate in and partake more richly of the life sciences at Cornell.

Future plans

As soon as VIVO was made public the Library began to receive inquiries about applying the technology to other disciplines and applications. As the tools underlying VIVO have gained maturity, we have endeavored to align VIVO more closely with external standards to permit data in VIVO to be explored using other semantic web tools, and to leverage other tools to bring new capabilities to VIVO [4].

The most exciting new development lies in the area of inferencing – leveraging the knowledge base of VIVO to discover data relationships not directly coded in the database. These additional relationships may form a basis for more advanced searching, browsing, and clustering capabilities. For example, we must currently explicitly identify faculty as members of a broad discipline such as life sciences or social sciences, since there is no set of criteria defining membership or affiliation. If instead we could identify participation in a discipline by the presence or absence of relationships including department or graduate field affiliations, and perhaps additionally the subject areas or venues of recent publications, the membership in this derived type could become a very useful interactive search limitation or a filter on display of data on a remote website syndicating VIVO content. Membership in the new, defined type could be evaluated and maintained automatically, much as the Lucene full-text search index is updated, via listeners that trigger a reclassification of any data added or changed through the editing interface.

As we have gained experience working with ontologies in practice, it has become more evident that no single, self-contained ontology will meet all our needs. By including name space identification in our ontology types and properties, we can effectively maintain distinct low-level ontologies mapping directly to each source for data as it arrives from an external database – as with our example before, differentiating clinical and teaching faculty. A middle-level VIVO ontology combines data to eliminate distinctions relevant in local sources but not across the whole university, as for example to report on "faculty" regardless of their type of appointment. Adding an additional layer of dynamically derived types will allow mapping from core characteristics of data into useful output aggregations such as international researchers using the presence of, or distinct values for, specific object or data properties.

Building defined types based on inferences requires that VIVO align itself very closely to semantic web standards and leverage tools such as Pellet (http://pellet.owldl.com/) and D2R (http://sites.wiwiss.fu-berlin.de/suhl/bizer/d2rmap/D2Rmap.htm) that have only recently become suitable for production use. Taken together, VIVO and the ontology mapping and reasoning capabilities offer significant immediate, practical applications of Semantic Web technology for libraries and other institutions seeking to manage complex information. The potential for classifying information as it arrives based on direct knowledge, while building up richer metadata stores via relationships developed and augmented over time, offers a fertile middle ground for exploration between manual data creation and automated data processing.

We expect to explore the use of this set of technologies as an integrating indexing tool for institutional and disciplinary research data repositories, managing metadata about multiple collections independently from any single repository but with rich cross-relationships supporting the interactive discovery of and navigation to content that may reside locally or across the world. Further, we hope soon to look beyond university boundaries to initiate and sustain collaborations with a handful of other universities, using this tool to create a larger inter-disciplinary academic community. A virtual community such as VIVO could well serve as a model to explore synergies with peer institutions, museums, foundations and research consortia to provide access to academic information on a national scale.

Acknowledgements

References

1. *Enhancing Science and Engineering at Harvard*. University Planning Committee for Science and Engineering. December 2006. Harvard University, Cambridge, MA. <http://www.provost.harvard.edu/reports/UPCSE_Interim_Report.pdf>.

2. *The University of Minnesota President's Interdisciplinary Initiative on Arts and Humanities*. Steven J. Rosenstone. July 15, 2004. College of Liberal Arts, Twin Cities Campus, Minneapolis, MN. <http://www1.umn.edu/pres/00_images/pdf/Abstract_7-15-04.pdf>.

3. *A Framework for Representing Knowledge*, Marvin Minsky, MIT-AI Laboratory Memo 306, June 1974. <http://hdl.handle.net/1721.1/6089>.

4. *A Semantic Web Primer for Object-Oriented Software Developers*. Holger Knoblauch, Daniel Oberle, Phil Tetlow, Evan Wallace. March 9, 2006. W3C Working Group Note. <http://www.w3.org/TR/sw-oosd-primer>.

5. *Ontology-Based Knowledge Representation for Bioinformatics*. Robert Stevens, Carole A. Goble and Sean Bechhofer. November 2000. Briefings in Bioinformatics 1 (4): 398 - 414.

Top | **Contents**
Search | **Author Index** | **Title Index** | **Back Issues**
Previous Article | **Next Article**
Home | **E-mail the Editor**

D-Lib Magazine Access Terms and Conditions

doi:10.1045/july2007-devare